DOCUMENT RESUME

ED 410 759 FL 024 702

AUTHOR	Thornton, Julie A.
TITLE	The Unified Language Testing Plan: Speaking Proficiency Test. Spanish and English Pilot Validation Studies. Report Number 1.
INSTITUTION	Center for the Advancement of Language Learning, Arlington, VA.
PUB DATE	1996-02-00
NOTE	243p.; For report #2, see FL 024 703.
AVAILABLE FROM	The Foreign Language Testing Board (FLTB) is an interagency group including representatives from the CIA, DIA, DLI, FBI, Department of State, and NSA.
PUB TYPE	Reports - Descriptive (141) -- Reports - Evaluative (142)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	*English; Federal Programs; *Language Proficiency; *Language Tests; Oral Language; Program Descriptions; *Spanish; Speech Skills; Standardized Tests; Standards; Test Validity; Testing; *Verbal Tests
IDENTIFIERS	*Federal Language Testing Board; *Unified Language Testing Plan

ABSTRACT

This report describes one segment of the Federal Language Testing Board's Unified Language Testing Plan (ULTP), the validation of speaking proficiency tests in Spanish and English. The ULTP is a project to increase standardization of foreign language proficiency measurement and promote sharing of resources among testing programs in the federal government. Over 200 individuals were tested in the two pilot studies. Analysis of results indicates higher reliability of the ratings than those of a prior interagency study, and increased reliability of the Spanish test than found in an earlier pilot study. Recommendations are made for pilot implementation, further development of tests in other languages, maintenance of interagency collaboration, coordination and quality control during pilot and full implementation, and adjustment in tester training workshop format. Appended materials include the examinee instructions, pre- and post-test questionnaires, rating frequency charts, a summary of Spanish results, a summary of English results, and cross-tabulation charts for both studies. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 410 759

DCI Director of Central Intelligence
Foreign Language Committee

Report Number 1

The Unified Language Testing Plan: Speaking Proficiency Test

Spanish and English Pilot Validation Studies

February 1996

BEST COPY AVAILABLE

FL024702

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Julie Thornton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Director of Central Intelligence
Foreign Language Committee

Report Number 1

The Unified Language Testing Plan: Speaking Proficiency Test

Spanish and English Pilot Validation Studies

February 1996

This paper was prepared by Julie A. Thornton under the direction of the Federal Language Testing Board at the Center for the Advancement of Language Learning.

CALL
Center for the Advancement
of Language Learning

2 February 1996

MEMORANDUM FOR: Susan Rudy, Chairman
DCI Foreign Language Committee

FROM: Betty A. Kilgore
Director

SUBJECT: Report #1 on the Unified Language Testing Plan:
Spanish & English Speaking Proficiency Test
Pilot Validation Studies

1. The document referenced above is the first report on an interagency test of foreign language speaking proficiency under the Unified Language Testing Plan (ULTP) of the DCI Foreign Language Committee. The success of the work detailed in this report has been made possible by the cooperation and assistance of many people and organizations. I would like to take this opportunity to acknowledge their efforts, without which the new speaking proficiency test format could not have been developed and piloted.

2. The Federal Language Testing Board (FLTB) includes representatives from the Central Intelligence Agency Office of Training and Education Language Training Division, Defense Intelligence Agency, Defense Language Institute Foreign Language Center Directorate of Evaluation and Standardization, Federal Bureau of Investigation Language Program, Department of State Foreign Service Institute School of Language Studies, and National Security Agency National Cryptologic School. The FLTB is the interagency working group that developed the ULTP and planned and implemented the first stages of the Plan under the direction of the FLC. FLTB members have worked intensively over the last 36 months to develop and validate the new Speaking Proficiency Test (SPT) format.

3. During the planning stage as well as during both pilot validation studies, agency representatives devoted considerable professional and personal time to interagency meetings, test development, materials and syllabus development, and the development and revision of this report as well as other documents related to the pilot validation project. Agency

representatives who participated in the process include Thea Bruhn, Marijke I. Cascallar, James R. Child, John L.D. Clark, Madeline Ehrman, Michael Furlo, Katrin Gardner, Dariush Hooshmand, Frederick H. (Rick) Jackson, Angela Kellogg, Anna Knight, Yvonne March, Olga Navarette, Suzanne Olson, Sigrun Rockmaker, and Burt Weisman. Pardee Lowe and Stephen Soudakoff participated in FLTB working meetings in an ex officio capacity as chair/moderator of the ILR Testing Committee.

4. Marianne Armstrong, Anne-Marie Carnemark, Marijke Cascallar, Raul Cucalón, Marisa Curran, Patricia Dege, Angela Kellogg, Yvonne March, Jose Molina, Sietske Semakis, Yakov Shadyavichyus, and Don Smith as well as other agency representatives on the Board participated in the critical role of tester trainers. I would like to thank them for their hard work.

5. The DCI Foreign Language Committee and the CALL Executive Committee invested much time and many resources in ULTP activities. I thank them for their consistent support.

6. I also express our appreciation to the management personnel, testing program managers, language teachers and linguists of the Central Intelligence Agency Office of Training & Education Language Training Division, the Defense Language Institute Foreign Language Center and English Language Center, the Language Program at the Federal Bureau of Investigation, the Foreign Service Institute at the National Foreign Affairs Training Center, and the Department of Justice Executive Offices of Immigration Review, who participated in the studies or who permitted their personnel to participate. I am also grateful for the support provided by many of the Interagency Language Roundtable (ILR) member organizations in providing examinees for the studies. I recognize that their participation often required a sacrifice of other duties. I recognize the excellent work each tester performed, and I am grateful to each of them for always giving their best effort. I would also like to acknowledge the cooperation of approximately 300 volunteers who agreed to participate in the study. I appreciate their goodwill and frank comments about the test, which have been invaluable to the test development process.

7. I also recognize the hard work and dedication of the CALL Testing staff, including Eduardo Cascallar, David Fielder, Shirley Parker, Claressa Strawn, Julie Thornton, and Alexandra Woodford. Eduardo Cascallar, the CALL Testing Coordinator, led FLTB test development and pilot planning discussions, designed the pilot validation studies, and provided expert technical advice to the FLTB. Julie Thornton, the Assistant Testing Coordiñator, assisted the FLTB throughout the process and served as coordinator and author of this report.

8. Lastly, I thank Prof. Fred Davidson from the University of Illinois, Champaign-Urbana for his assistance in developing and reviewing the report. His comments and suggestions were invaluable. I appreciate in particular his expert perspective on testing issues central to the report.

Betty A. Kilgore

Betty A. Kilgore

Contents

Page

Section 1. Executive Summary 1
Section 2. Introduction 5
  Interagency Task Force 5
  Federal Language Testing Board 6
  Unified Language Testing Plan 7
  Timeline for the Unified Language Testing Plan 7
  Modifications to the Unified Language Testing Plan 7
Section 3. Test Description 11
  Speaking Proficiency Test Objective 11
  The ILR Criterion 11
  Test Format 12
    Test Phases 12
    Elicitation 13
  Rating 15
    Rating Factor Grid 16
    Rating Factor Definitions 16
    The Rating Process 18
Section 4. Test and Training Materials Development 19
  Test Specifications 20
  Spanish Participant's Packet 20
  English Tester Manual 20
  Instructions 20
  Additional Materials 21
Section 5. Pilot Study Research Design 23
  Testers 23
    Spanish 23
    English 24
  Subjects 25
    Spanish 26
    English 26
  Instructions 27
  Examinee Questionnaires 27
  Tester Questionnaires 28
  Data Collection Procedures 28
    Spanish 28
    English 29
Section 6. Rating Reliability Results 31
  Use of the ILR Scale 31

Frequency Charts: Spanish and English Pilot Studies 31
Normality Results: Spanish and English Pilot Studies 31
Note on Measures Used in Reliability Section 32
Percent Level of Agreement 32
Non-Parametric Statistical Analyses 33
Spanish Pilot Study 36
Interagency Reliability 36
Intra-Agency Reliability 44
Inter-Rater Reliability 45
Effects on Reliability by Test Order/Time of Administration 47
English Pilot Study 50
Inter-Pair Reliability 51
Inter-Rater Reliability 57
Effects on Reliability by Test Order/ Time of Administration 58
Effect on Reliability by Location of Training 60
Section 7. SPT Validity 63
Current Thought on the Notion of Validity 63
Literature Review: A Unified Concept of Validity 63
Convergent Evidence of Validity 64
Concurrent Evidence of Validity 64
Face Evidence of Validity 65
Examinee Feedback 65
Tester Feedback 66
Content Evidence of Validity 67
Validity: Concluding Remarks 69
Section 8. Recommendations 71
Section 9. Bibliography 75

Appendixes

A. Examinee Instructions A-1
B. Pre-Test Questionnaire B-1
C. Post-Test Questionnaires C-1
D. Frequency Charts D-1
E. Summary Spanish Results E-1
F. Summary English Results F-1
G. Crosstab Charts for the SPT Spanish Study G-1
H. Crosstab Charts for the SPT English Study H-1

5

Section 1. Executive Summary

The Federal Language Testing Board (FLTB) of the Center for the Advancement of Language Learning (CALL) has been tasked with developing and implementing the Unified Language Testing Plan (ULTP). The ULTP was established in 1994 as a part of the National Performance Review headed by Vice President Albert Gore. The main objectives of the ULTP are to increase the overall standardization of foreign language proficiency measurement and to promote sharing of resources among testing programs in the Federal government. The ULTP provides for general proficiency assessment of speaking, listening, reading, and writing. The FLTB chose the measurement of speaking proficiency as its first area of focus. As of November 1995, pilot studies have been completed on the Speaking Proficiency Test (SPT) in two languages (Spanish and English as a Second Language). A third pilot study in Russian is nearing completion. This document reports on the development of the test and the results of the first two studies (Spanish and English).

Test Development

The following are specific accomplishments under the ULTP:

- *Test Specifications* developed and agreed to by the FLTB representatives
- Interagency training syllabus and materials developed
- Computer-delivered self-study multimedia training program launched and piloted
- *Tester Manual* developed
- Interagency groups of Spanish and English testers trained and qualified to test
- Over 200 subjects tested in Spanish and English pilot studies by testers from the Central Intelligence Agency, Defense Language Institute, Federal Bureau of Investigation, and Foreign Service Institute in the Spanish pilot study and by testers from Defense Language Institute Foreign Language Center in Monterey, Defense Language Institute English Language Center in San Antonio, Federal Bureau of Investigation, and the Department of Justice Executive Office of Immigration Review in the English pilot study.

Results

Analysis of results from both Speaking Proficiency Test pilot validation projects indicates higher reliability of ratings than those of the only prior interagency study, performed in 1986 by the Center for Applied Linguistics, where three agencies (CIA, DLI, and FSI) administered tests according to each agency's testing procedures in place at that time to the same set of examinees. A further indication of progress is that a number of results from the second SPT pilot study (in English) show increased reliability over those of the first pilot study conducted in Spanish.

The following research questions addressed areas of particular importance drawn from the Spanish and English studies. Additional detail is provided in the Results section of this report.

Research Question #1: If a given examinee, after receiving his or her score based on the standard two-member testing pair procedures, complains about or questions the results, how likely is it that the original score would be duplicated if the examinee were to be tested and rated by a second (randomly selected) testing pair? (An exact match requires that both pairs agree exactly. A within-level score match requires that the ratings fall within the same base level, e.g., a 2 and a 2+.)

	within-level matches	**exact matches**
English (1995)	57%	42%
Spanish (1994-95)	57%	37%
French (1986)	47%	30%
German (1986)	41%	26%

Research Question #2: If an examinee was given four tests, each with different testers, what percentage of the time was he or she given exactly the same score in every test?

English (1995)	17%
Spanish (1994-95)	12%

Research Question #3: What percent of the examinees tested in each study received a *different* score in each of their tests?

English (1995)	1% (four tests each)
Spanish (1994-95)	5% (four tests each)
French (1986)	30% (three tests each)
German (1986)	33% (three tests each)

Research Question #4: What happened to the scores if half of the testers are trained in one location and the other half are trained in another?

In the English study, four testers were trained on the west coast at DLI Monterey and four on the east coast at CALL. All trainees had no prior experience as language testers. The comparison of ratings of east coast trained testers with those of west coast trained testers resulted in a percent level of agreement of 42%. Within their testing pairs, individual testers agreed on their final ratings an average of 68% of the time: the east coast training group agreed with their respective testing partners 53% of the time, and the west coast group for 76% of the time—both very acceptable inter-rater reliability levels for novice testers.

Research Question #5: When two testers administered and scored the same SPT, how well did their initial individual ratings agree?

The average inter-rater agreement was 84% in the Spanish study. The testers in each pair were more different from one another in the first half of the study (percent agreement was 79%) and tended to be more similar in the second half (89% agreement). Each tester worked with the same testing partner for the entire study; these results indicate that the Spanish testers grew more similar in rating over time, even as they seemed to drift slightly further apart in rating from the other pairs. The average inter-rater agreement for the novice testers in the English study was 68%. The English testers showed the opposite trend, agreeing with one another more in the first half (70%) and less in the second half (66%).

Research Question #6: What did the testers and examinees think of the new SPT?

Each tester and each examinee who participated in both pilot studies was asked to provide detailed feedback on their experiences. Both tester and examinee feedback on the new test was supportive and highly encouraging.

Recommendations

- Based on the positive results from the Spanish and English studies and preliminary results from the Russian study, begin moving at the various agencies toward pilot operational implementation, resources permitting.
- Contingent upon a positive outcome to the Russian pilot study and pilot operational implementation projects and upon individual agency approval, fully implement the SPT in all languages.
- Maintain ongoing interagency collaboration on language proficiency testing.
- Coordinate interagency work on pilot operational implementation projects and quality control procedures during pilot and full implementation.
- Consider adjusting the format of SPT tester training workshops, based on results of these studies that indicate that retraining of experienced testers requires more time than originally expected. Additional time may be dedicated to formal classroom learning, practice testing, and apprenticeships in addition to possible self-study components.

Conclusion

The process of development and pilot testing of the SPT has produced a test that meets many of the oral proficiency testing needs of participating government agencies with a higher reliability of scores than previously demonstrated in interagency testing. The development of the SPT has further encouraged, and benefited from, an increased level of interagency cooperation and sharing of resources. With the implementation of the interagency SPT, resources can continue to be shared to provide for more efficient and cost-effective testing programs, and test results will be meaningful and exchangeable across agencies.

Section 2. Introduction

This document reports on work in progress under the Unified Language Testing Plan (ULTP), which was developed for the Director of Central Intelligence Foreign Language Committee (DCI/FLC) by the Federal Language Testing Board (FLTB) at the Center for the Advancement of Language Learning (CALL). It specifically addresses the development of the Speaking Proficiency Test (SPT) and the initial validation studies performed on this test.

The FLTB consists of representatives from the following federal agencies:

- Central Intelligence Agency (CIA)
- Defense Intelligence Agency (DIA)
- Defense Language Institute (DLI)
- Federal Bureau of Investigation (FBI)
- Department of State, Foreign Service Institute (FSI)
- National Security Agency (NSA)

CALL provides professional guidance and consultation as well as administrative support for FLTB activities. The moderator of the Interagency Language Roundtable (ILR) Testing Committee participates in all FLTB meetings in an ex officio capacity.

The ULTP was developed and approved in February 1994 in response to the National Performance Review recommendation for the setting of "*. . . Community-wide language proficiency standards. . . .*" It provides a single, long-term plan to integrate the government's language testing system while at the same time accommodating the job-related language testing needs of each participating agency.

The ULTP was designed by the FLTB to satisfy the need for a common, interagency, general proficiency assessment of speaking, listening, reading, and writing. The approach chosen addresses this need through a multi-year program, which starts with the development, piloting, and implementation of a common oral proficiency test and continues with the development, in turn, of common testing procedures for listening, reading, and writing. The approach is rigorous in ensuring that each new test demonstrate acceptable validity and reliability before full implementation.

Interagency Task Force

Beginning in 1992, when funding was set aside to create the Center for the Advancement of Language Learning (CALL) under the FY 1992/93 Foreign Language Initiative, it was determined that one area of focus for CALL would be testing, to be coordinated by an interagency testing board. Once the goals for the testing area had been established, CALL set up an interagency task force to create a plan to achieve those goals. Representatives from each of the four USG language schools met at CALL for a five-week assignment. These representatives were working language teachers and testers familiar with their agencies' current testing practices and needs. Drawing upon their experience and expertise in language proficiency testing, the task force scrutinized the language proficiency definitions used by the Community. The result of their work included a set of recommendations for the creation and initial operation of a Language Proficiency Testing Board. The task force concentrated first on oral proficiency testing, expecting that the review of speaking assessment would serve as a model for later consideration of tests of listening and reading proficiency.

The task force concluded in its report (Armstrong, et al., 1992) that the current system of oral proficiency testing in some cases does not meet the government's overall needs because it is tailored to the specific needs of the individual agencies. The four language schools had focused more on the missions of their respective agencies than on collective efforts to address Community-wide requirements. Although there was significant collaboration among the agencies, that collaboration was generally not conducted in a coordinated or systematic way. The task force proposed the creation of a uniform proficiency testing system. The task force members reported that a uniform system of this type would require resources to develop a theoretical testing model, a standardized format, uniform testing procedures, uniform scoring procedures, and provision for quality control within the system. The task force identified the steps necessary to create a uniform testing system and planned an organizational structure, the Advisory Panel of the Language Proficiency Testing Board to perform those activities. The task force submitted its recommendations to the CALL participating agencies, and the Language Proficiency Testing Board (later renamed the Federal Language Testing Board to reflect more accurately the scope of its mission) was created.

Federal Language Testing Board

The Federal Language Testing Board began meeting at the beginning of January 1993 with Dr. John Clark of DLI as interim Testing Coordinator. This panel was made up of testing program managers from the six agencies that participate in CALL's Executive Committee (CIA, DLI, DIA, FBI, FSI, and NSA) as well as the moderator from the ILR Testing Committee (as a non-voting member). Regular meetings were held by the Panel throughout 1993, during which time the members discussed issues related to speaking proficiency testing and began to build consensus about the characteristics of an interagency speaking test. Dr. Eduardo Cascallar was named the FLTB coordinator in March 1993 and served in this position until the end of September 1995. During this time, the FLTB also changed the name of the panel to the Federal Language Testing Board. During these early discussions, participants developed a greater understanding of each agency's testing needs, specific testing methods, identified areas of similarity and difference in those methods, and became better acquainted with their colleagues from the other agencies. Various approaches to a plan for a unified language testing system were explored and developed.

Unified Language Testing Plan

The DCI Foreign Language Committee gave the FLTB the task of creating a plan to respond to the National Performance Review recommendation to the Intelligence Community for the setting of ". . . Community-wide language proficiency standards . . . ." In early 1994, the FLTB developed the Unified Language Testing Plan (ULTP). The ULTP was approved by the Foreign Language Committee in February 1994 and published in March 1994. (Copies of the ULTP are available upon request from CALL.) The ULTP includes a timeline for the development, validation, and implementation of a new interagency test of speaking proficiency, as well as later projects to address the other skills of listening, reading, and writing. This timeline focuses first on speaking test development and charts the development of a clear set of test specifications for the interagency format, three pilot validation studies, and implementation of the new test format across all agencies. The ULTP calls for the new SPT format and procedures to be validated in three languages. The languages originally chosen by the Foreign Language Committee for these studies were Spanish, Russian, and Chinese. These languages were selected on the basis of two criteria: (1) that all participating agencies could provide testers and examinees in these languages and (2) that, by performing validation studies in these languages, the test would have been validated in languages of different levels of difficulty for native speakers of English.

Timeline for the Unified Language Testing Plan

Some changes have been required in carrying out the details of the planned ULTP timeline; however, most of the substantive work of the FLTB has proceeded on schedule. Originally the development and piloting of the SPT were scheduled to be conducted in Spanish (from November 1994 to February 1995), followed by Russian (February to May 1995), and then Chinese (May to August 1995). Operational implementation of the SPT at the agencies was planned to begin in early 1996. Working sessions for the development of a new listening proficiency test were scheduled to begin in June 1995 with a similar project to begin on reading in July 1996.

Modifications to the Unified Language Testing Plan

Four significant modifications have been made to the ULTP timeline:

(1) Due to unexpected resource constraints in Russian language training for the 1994/95 academic year, participating United States Government (USG) language schools were unable to release the Russian faculty needed to participate in the planned Russian tester training and pilot testing in the spring of 1995. The Russian language tester training and pilot study were rescheduled for the summer of 1995.

(2) To replace the postponed Russian study in the spring, a smaller scale, empirical study was conducted using English as a Second Language (ESL) as the test language. The FLTB felt that this study would uniquely provide materials that could be used in future tester training for all languages. Furthermore, the original research design was altered for the English study so that (a) the training took place at two different sites with two completely different training teams and (b)

the training was provided to individuals with no experience as testers. These alterations enabled the FLTB to increase the information obtained from the study.

(3) Because the preliminary results from the Spanish and English studies have been encouraging, the FLTB has decided that it may be possible for individual agencies to begin planning pilot operational implementation projects of the SPT training and testing procedures in fall 1995. These pilot projects will be possible provided that the results of the Russian study prove favorable. Two agencies have already made plans to go ahead with such projects; other agencies will decide after they have seen the final results and determined what resources are available for this activity. Pilot implementations are expected to be conducted in various languages, including Spanish, Russian, and/or English, and will entail reports to the FLTB and rigorous study of interagency reliability by having an appropriately drawn sample of tests also rated by other agencies. If the results of both the Russian pilot study and the pilot operational implementation at the respective agencies are positive, the agencies might be able to begin full operational implementation of the test by summer 1996, which would be ahead of the original ULTP schedule.

(4) Work on the development of an interagency test of listening proficiency, originally scheduled to begin in June 1995, has been postponed until the Testing Committee of the Interagency Language Roundtable (ILR) has completed its review and revision of the ILR Listening Skill Level Descriptions, which will almost certainly provide the foundation for future listening test development and scoring. The ILR Testing Committee will reconvene in early 1996, at which time it should be possible to estimate when the revision of the Guidelines will have proceeded far enough to permit the FLTB to move forward with test development.

Unified Language Testing Plan **Accomplishments and Projected Timeline**	
FY93/94	• Unified Language Testing Plan developed, approved by the Foreign Language Committee, and published (March 1994) • FLTB working sessions on **Speaking** test specifications, tester training curriculum design, and materials development (January 1994 to September 1994)
FY94/95	• Spanish tester retraining (October 1994) • Spanish pilot testing (November 1994 to February 1995) • Revisions to the test based on Spanish results (January 1995 to April 1995) • Spanish statistical analysis (beginning in December 1994) • English tester training (April 1995) • English pilot testing (May to June 1995) • English statistical analysis (beginning in June 1995) • Revisions to the test based on English results (July 1995) • Preliminary status report published (August 1995) • Russian tester retraining (July 1995) • Russian practicum/formative phase (July to August 1995) • Russian pilot testing (September 1995 to November 1995)

Unified Language Testing Plan **Accomplishments and Projected Timeline** (continued)	
FY95/96	- Begin pilot operational implementation of SPT (First Quarter 1996)
	- Final report on Spanish and English published (February 1996)
	- Begin FLTB working sessions on **Listening** (February 1996)
	- Final report on Russian study (May 1996)
	- Final combined report—all studies (July 1996)
	- Begin SPT reliability/retraining program (August 1996)
	- Begin implementation in all languages (September 1996)
FY96/97	- Begin FLTB working sessions on **Reading** (December 1996)
	- Continue SPT implementation in all languages
FY97/98	- Begin FLTB working sessions on **Writing** (December 1997)

Section 3. Test Description

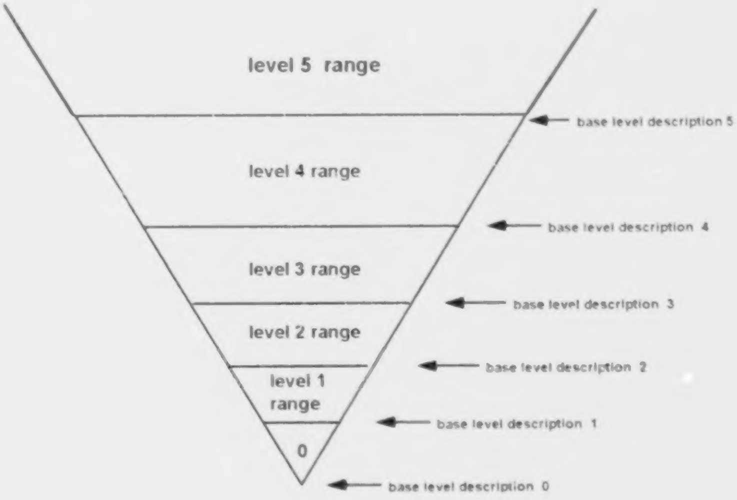
The following section describes the SPT test objective, the rating criteria, the test format, SPT elicitation techniques, and SPT rating procedures.

Speaking Proficiency Test Objective

The goal of the SPT is to have testers elicit, or obtain, a sample of an examinee's speech performance that can be matched reliably to an appropriate ILR Speaking Skill Level Description. The firmly established ILR descriptions, which range from "Level 0–No Proficiency" to "Level 5–Functionally Native Proficiency," are the final rating criteria for the SPT. Testers use specific techniques to elicit the needed language sample from the examinee. The objective of this elicitation process is to ensure that the sample is, in fact, indicative of the examinee's true ability and that it will be ratable according to the ILR descriptions. Final rating takes place immediately following the test, after the full speech sample has been obtained.

The ILR Criterion

The ILR Speaking Skill Level Descriptions characterize a full range of speaking proficiency. The complete ILR scale is divided into six base levels (0 to 5), each of which, in itself, represents a range of proficiency. These ranges do not appear at regular intervals on the overall scale nor do they represent equal amounts of language proficiency. Rather, the ILR levels increase in size progressively such that the scope of additional functions and tasks controlled at level 1, for example, is much smaller than that controlled at level 2. Each level also includes the language abilities described by all lower levels.


level 5 range
base level description 5
level 4 range
base level description 4
level 3 range
base level description 3
level 2 range
base level description 2
level 1 range
base level description 1
0
base level description 0

BEST COPY AVAILABLE

The descriptions for each level indicate minimum performance requirements for that level. The upper range of ability for a given level will go substantially beyond the base level description, but it will not consistently meet the requirements of the next base level. The base level descriptions are considered thresholds in that the proficiency requirements that they describe must be completely met for an examinee to be placed within that range. Because the ranges are broad, two examinees receiving the same rating may actually exhibit different strengths and weaknesses in the test language. What they will have in common, however, is the ability to fulfill all of the minimum requirements of the level at which they are rated and the inability to meet all of the threshold requirements for the next base level.

In addition to the base levels, the ILR also describes five "plus" levels (0+ through 4+). The plus levels are not considered thresholds; they fall within the level ranges delineated by the base levels. Plus-level descriptions indicate proficiency that "substantially exceeds one base skill level and does not fully meet the criteria for the next base level." Base levels and plus levels are treated differently during rating in the SPT. (The rating process is described below under *Rating*.)

Test Format

The SPT is a face-to-face interactive test in which two trained testers speak with an examinee on a variety of topics for approximately 15 to 45 minutes. Ideally, the testers would both be educated native speakers of the test language, speakers of English at the professional level, and trained and certified testers in the test language. In cases where it is operationally impossible to meet these criteria, one of the testers may be less than fully equivalent to an educated native speaker of the test language, and/or one of the testers may have only elementary proficiency in English.

Under normal circumstances, both testers interact with the examinee in a three-way conversation. In addition to conversation, other activities are included in the SPT. These activities will be more fully described under *Elicitation*. To assign roles for the presentation of these activities and to select possible topic areas for inclusion in the test, the two testers are required to meet before the start of the test for a brief pre-planning session.

The examinee enters the testing room and is greeted by the testers. One of the testers provides oral instructions to the examinee in English. These instructions reiterate the major points detailed in the written "Instructions for the Examinee" sheet, which each examinee receives before entering the testing room. Once the examinee indicates an understanding of the test instructions, the testers begin to interact with the examinee in the test language.

Test Phases

Each SPT consists of three phases: the Warm-Up, the Core of the Test (consisting of iterative level checks and probes), and the Wind-Down.

Warm-Up. The purpose of the Warm-Up in each test is to put the examinee at ease and to give the testers an initial indication of the examinee's proficiency level. The Warm-Up consists of fairly simple, polite, informal conversation. The Warm-Up generally lasts from one to three minutes, the length depending on the apparent readiness of the examinee to be challenged in the next phase. The Warm-Up will usually be longer for lower-level examinees.

Core of the Test. The Core of the Test is the main body of the Speaking Proficiency Test. The purpose of the Core of the Test is to find the examinee's level of sustained ability in the test language as well as the limits of that ability. The key activities performed in this phase are described under *Elicitation Activities*.

Wind-Down. The purpose of the Wind-Down is to ensure that the examinee leaves the test with a feeling of accomplishment. The Wind-Down consists of brief, informal conversation on a topic comfortable for the examinee followed by appropriate leave-taking. The language level used should be comfortable for the examinee and should not challenge him or her. At the same time, the Wind-Down should not be conducted at an inappropriately low level.

Elicitation

Elicitation refers to the activities undertaken by testers within a test to draw a ratable language sample from the examinee.

Definition. To establish evidence of the examinee's strengths and weaknesses in the test language and to obtain a sufficiently broad sample of speech for rating, SPT testers are required to elicit the following elements from an examinee:

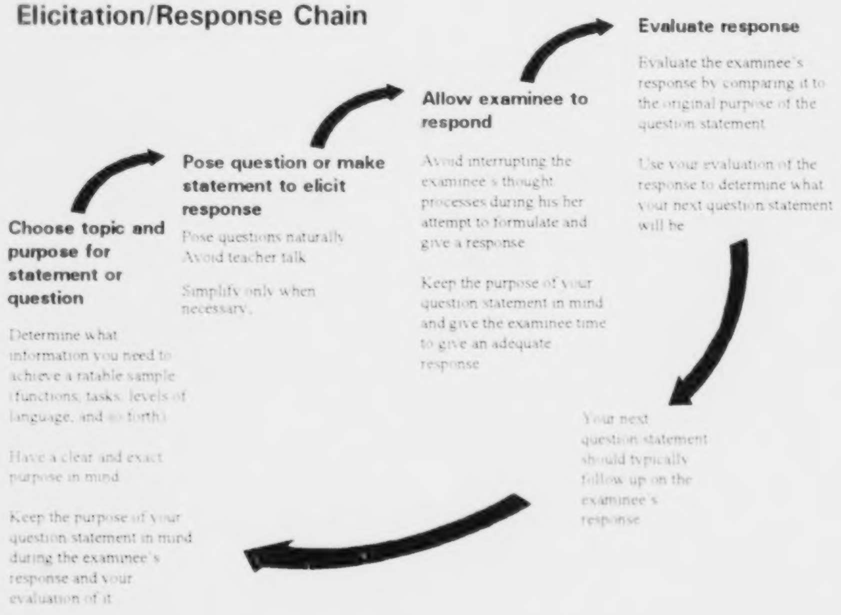
- Samples of interactive conversation.
- Multiple language functions and tasks.
- Multiple topics.
- Samples of examinee eliciting information from a tester and demonstrating comprehension.
- Samples of extended speech on a topic with little or no interruption.
- Instances of language breakdown.

While covering these required areas during the Core of the Test, testers also must continuously verify the maximum level of speaking proficiency the examinee can sustain. This process, called *level checking*, establishes the *working level*, the level which testers hypothesize, up to any given point in the test, to be the actual proficiency level of the examinee. At the same time, testers need to collect evidence that the examinee cannot sustain performance at any higher level. The process of pushing the examinee to the point where his or her language is insufficient is called *probing*. During the Spanish and English studies, each test was to contain at least two failed probes. The object of probing is to find points of *language breakdown*, defined as any time in the test at which the examinee is unable to accomplish a language task in a manner that satisfies the performance expectations of the level being probed. Adjustment of the working level is often necessary during a test; for example, when the examinee fails to sustain speech at the working level, the working level must be lowered; or when the examinee succeeds in performing tasks at the probe level, the working level must be raised.

Elicitation Activities. Carefully structured, purposeful conversation with the examinee is the primary activity in which the testers engage to accomplish their elicitation goals. Two other types of activity may be, and typically are, used to complement the conversational core of the test. These are known as Situations and the Information Gathering Tasks (IGT).

Conversation. The Core of the Test consists, for the most part, of conversation-based elicitation. Testers ask questions or make statements to engage the examinee in a conversation. Testers were introduced to a number of question/elicitation types during the SPT training workshops in both studies. This set is a subset of those question/elicitation types used in current tests at the various agencies. The range of conversation topics and tasks the testers introduce during this conversation serve to test the overall strengths and weaknesses of the examinee. Testers select questions or statements carefully so as to elicit aspects of speech that will enhance the sample and that are appropriate in light of the abilities demonstrated by the examinee to that point. The following Elicitation/Response Chain is used to illustrate the testers' process of focused questioning.

Elicitation/Response Chain



Choose topic and purpose for statement or question

Determine what information you need to achieve a ratable sample (functions, tasks, levels of language, and so forth).

Have a clear and exact purpose in mind.

Keep the purpose of your question/statement in mind during the examinee's response and your evaluation of it.

Pose question or make statement to elicit response

Pose questions naturally. Avoid teacher talk.

Simplify only when necessary.

Allow examinee to respond

Avoid interrupting the examinee's thought processes during his/her attempt to formulate and give a response.

Keep the purpose of your question/statement in mind and give the examinee time to give an adequate response.

Evaluate response

Evaluate the examinee's response by comparing it to the original purpose of the question/statement.

Use your evaluation of the response to determine what your next question/statement will be.

Your next question/statement should typically follow up on the examinee's response.

Situations. Situations, or role plays, place the examinee and one of the testers in an imaginary, test-culture setting where they act out a scenario. The examinee is asked to accomplish a specific task in an interaction with the tester. In each test, testers choose Situations that are realistic and appropriate for the examinee in level and in context. Situations are presented by one tester either in writing or, in some cases, orally and indicate the scene, the examinee's role and objective, and the tester's role. The examinee is never asked to play someone other than himself or herself.

Situations are used by testers to draw aspects of language use from the examinee that cannot be easily demonstrated otherwise. Situations are useful for testing the examinee's ability to use appropriate speech register when a particular relationship requires him or her to do so, to communicate effectively and appropriately in contexts other than polite informal conversation, and to interact appropriately with a native speaker in a test-culture setting. Situations can be used to elicit survival language, concrete language, register shift, vocabulary range, cultural aspects, or the ability to influence. In the SPT, Situations are not tied to a specific ILR level; instead, the testers select Situations that will improve the sample of speech obtained from the examinee.

There are two types of Situations: basic/routine and non-routine. In both types of Situations, the examinee performs tasks that might be required of someone using a foreign language while living and/or working abroad or when interacting with speakers of the test-culture. However, non-routine Situations are not predictable, everyday transactions. They may involve the need to solve a problem, to get out of a predicament, to try to influence someone to do something or to change an opinion, or to explain a special set of circumstances. Basic Situations can be made non-routine through the introduction of complications or linguistic or cultural complexities.

The Information Gathering Task (IGT). One way to give the examinee the opportunity to elicit and demonstrate comprehension of information from a tester (one of the requirements of a ratable sample) is to have the examinee perform an IGT.

One purpose of this task is to give the examinee the opportunity to elicit information from one of the testers and, in doing so, to show how well he or she can manage the interaction and gather information in the test language. Another important purpose is to give the examinee the opportunity to demonstrate his or her comprehension of the test language and the strategies he or she uses to verify understanding.

The IGT is generally introduced toward the end of the Core of the Test, usually after the Situation. One tester asks the examinee to interview the other tester on a specific topic. The examinee is given paper and pencil to take notes. The examinee interviews the tester in the test language. After three to five minutes, the examinee reports back to the first tester, typically in English, the information he or she elicited. After the report is finished, the testers may ask the examinee to provide additional clarification or explanation as needed to get a fuller sample.

Topics for the IGT may be anything about which the tester being interviewed feels comfortable speaking and that suits the interests and the language level of the examinee.

During the Spanish pilot study, both testers usually remained in the room during the IGT. This allowed both testers to hear all of the examinee's speech. During the English study, the tester who was not being interviewed left the room.

Rating

Rating is the process of determining the examinee's official ILR level score, based on the sample of speech elicited during the test. Testers (in their roles as raters) compare the elicited sample to the stated criteria of the ILR Skill Level Descriptions, which are the sole criteria for final rating. Raters verify that the examinee both consistently meets the stated requirements of the base level to be assigned and does not consistently meet the stated requirements of the next higher base level. Assigned ratings should correspond to the highest level at which the examinee performed consistently during the test.

Rating Factor Grid

A rating factor grid is used as a rating aid to help raters focus their assessment at appropriate ILR level ranges. However, an analysis of examinee performance on the rating factors alone does not produce a final rating.

The rating factor grid contains descriptions for six different rating factors separated according to ILR base levels. The majority of the statements contained in the factor grid are taken directly from the ILR descriptions. Some additional characteristics of the different factors were included by the Board. The six rating factors are:

- Interactive comprehension
- Structural control
- Lexical control
- Delivery
- Social/cultural appropriateness
- Communication strategies

Rating Factor Definitions

The following definitions were developed by the FLTB for each of the six rating factors and included in the October 1994 version of the *Test Specifications*. These definitions were the official rating factor definitions used in the Spanish and English pilot studies.

Interactive Comprehension. Refers to the ability of the examinee to comprehend the speech of a native speaker of the test language in conversation, where it is possible for the examinee to request clarification or amplification. Includes reference to whether the examinee is able to comprehend natural tester speech or requires the tester to produce slower and/or simplified speech and/or to adjust to the examinee's limitations in other ways. However, occasional requests for clarification do not in themselves indicate weaknesses in this skill factor. Comprehension is evidenced by the appropriateness with which the examinee responds to the tester and follows up on the tester's statements; it may also be evidenced by reporting what has been comprehended (either in English or in the test language). This factor includes general comprehension or gist but also includes comprehension of implicit and explicit structural relationships; lexical denotation and connotation; relationships signaled by register, nuance, irony, tone; and the pragmatics of utterances. At high levels, it also includes comprehension of cultural concepts quite different from the examinee's own, as well as of non-standard or regional dialects that would be generally understood by native speakers functioning at that level.

Structural Control. Refers to the accuracy and flexibility with which the examinee is able to use the language's morphological and syntactic structures to produce well-formed and appropriate sentences. Also refers to the examinee's ability to link sentences together appropriately in discourse to form longer utterances that are coherent and cohesive. Among the elements included within this factor are control of word order, grammatical markers such as those for tense, aspect, or complementation in some languages; derivational and inflectional affixes; modification; topicalization; and coordinate and subordinate conjunction. Structural control is evidenced by the well-formedness and cohesion of sentences and of connected discourse and by the range of different structures used by the examinee.

Lexical Control. Refers to the range and depth of vocabulary and idiomatic phrases on which the examinee is able to draw in speaking the language and the facility and appropriateness with which the examinee uses them. At upper levels, there is evidence of one or more professional vocabularies in addition to a broad, general one. May also refer to the use of proverbs, sayings, jokes, and other memorized scripts. Lexical control is evidenced through appropriateness and precision in selecting lexical items to achieve communicative purposes.

Delivery. Refers to the fluency and phonological accuracy with which the examinee produces utterances in the language. Fluency refers to the ease of flow and natural soundingness of the examinee's utterances. Phonological accuracy refers to the examinee's pronunciation in context of the individual sounds of the language and to the patterns of intonation, including stress and pitch. Delivery is evidenced by the extent to which utterances sound native-like, are smooth-flowing, and are free of features that interfere with communication of meaning.

Social/Cultural Appropriateness. Refers to the extent to which the examinee's use of the language is appropriate to the social and cultural context and reflects an understanding of cross-cultural communication. Includes control of body language and such paralinguistic elements as use of space-fillers to hold the floor in a conversation, back-channeling to indicate attention, and loudness or softness of speech, as well as selection of topics appropriate to the situation. Also includes control of several linguistic elements, including phatic scripts for occasions such as greeting, leave-taking, expressing condolences or congratulations, beginning or ending a story, or toasting; informal and formal registers; turn-taking conventions in a conversation; rhetorical devices and organization in connected speech; and culturally appropriate pragmatics. Evidence of social/cultural appropriateness is important at all proficiency levels but becomes crucial at the professional level (level 3) and beyond.

Communication Strategies. Refers to the examinee's ability to use discourse and compensation techniques to carry out communicative tasks. At lower and intermediate proficiency levels, these strategies typically take the form of compensating for weaknesses in comprehension or production by managing the interaction (taking control of the topic and/or the interaction where necessary) and by using such techniques as circumlocution, paraphrase, requesting clarification, and so forth. As proficiency levels rise, the range and sophistication of strategies available for repairing interactions increase. At upper proficiencies, this factor will frequently appear as the ability to plan and effectively carry out a complex communicative task and to negotiate meaning in ways that are nearly imperceptible, although they may be sometimes non-native.

The Rating Process

Considerable preliminary rating activity takes place during the test itself as the testers elicit a sample. Testers must form an initial working hypothesis of the examinee's proficiency early in the test and must continuously evaluate and modify this hypothesis during the test, based on the results of the probes and level-checks. However, no rating hypothesis is final until all necessary level-checks have been carried out successfully, the test has been concluded, and the following rating steps taken.

1. Each rater individually creates a preliminary profile using the rating factor grid to rate the examinee's performance on each of the six rating factors.
2. The performance profile from the rating factor grid completed in step 1 indicates the level at which the rater should begin to consult the ILR Speaking Skill Level Descriptions. The rater reads the ILR descriptions to determine the base level that fits the examinee's best consistent performance. The raters read only the level descriptions without the examples section. (If needed, a rater may consider the examples subsequently for further clarification, bearing in mind that the information in the examples section represents possible performances only.) The rater continues reading the descriptions of each successively higher base level until he or she identifies a base level for which the examinee has not met all the requirements. The rater assigns the next lower base level as the examinee's base rating, since this was the highest level for which all of the requirements were met.
3. To determine whether to assign a plus level rating, each rater rereads the description of the assigned ILR base level and its corresponding "plus" level. He or she decides which of the two descriptions best matches the examinee's performance. The rater then assigns this level as his or her individual final rating, noting observed strengths and breakdowns.
4. Then the raters negotiate a final rating for the test. As they negotiate this final rating, they discuss the test and their reasons for assigning their individual final ratings, and they review their perceptions of the examinee's performance during the different elicitation activities in the test to resolve any differences in their assessments.

In cases when the testers do not agree after negotiating, the test is marked as discrepant and sent to a third rater to resolve the discrepancy.

Section 4. Test and Training Materials Development

The section below will describe the materials developed in preparation for and refined during the Spanish and English validation studies.

The following is a general time frame for SPT materials development in preparation for the Spanish and English pilot studies:

- Fall 1994
  - — *Test Specifications* document finalized for use in Spanish pilot study.
  - — Training syllabus developed.
  - — Participant packet, containing guidelines for administering and scoring the SPT, developed for use in Spanish tester training.
  - — Videotapes of sample SPTs conducted by FLTB members and trainers produced in Spanish and English.
  - — Existing agency *Situations* reviewed and selected by FLTB.
  - — Draft examinee instruction script and sheet developed.
  - — Feedback questionnaires developed for examinees and testers.
- Spring 1995
  - — SPT *Tester Manual* created and revised for use in English tester training.
  - — Tester training syllabus revised extensively for use in English workshops.
  - — Videotapes of English sample tests created by trainers.
  - — Additional high-level Situations created, other *Situations* revised.
  - — Examinee instructions revised.

In preparation for both Spanish and English tester training workshops, the FLTB and tester trainers from the participating agencies met extensively to develop and review training syllabi and materials. These sessions-further provided the opportunity for introducing the SPT procedures to new trainers and receiving their feedback.

Test Specifications

In October 1994, just before the Spanish tester training workshop, a set of interagency, FLTB-approved test specifications was prepared. These specifications represent a set of principles for the development of the new test and outline the basic format of the proposed testing procedure. These specifications, drafted in the fall of 1994, have served as the basis for development of all aspects of the SPT and its related training materials.

Spanish Participant's Packet

Before the Spanish tester training, a participant's packet of study materials was assembled. This packet contained the test specifications, test administration information, a section on elicitation techniques, and a set of Situations. This packet was supplemented by handouts provided to testers during the training.

English Tester Manual

The manual that the English testers received at the beginning of their training was considerably different from the packet that the Spanish trainees received. The new manual was based on and included much of the same information as the Spanish participant's packet, but it was expanded and revised extensively by the FLTB between the end of the Spanish pilot and the beginning of the English tester training workshop. FLTB members provided extensive input to the development of the manual's contents and organization. The manual layout was completely revised to include more illustrations and white space, improving its readability and making it easier to follow. A number of charts and graphs were added to the body of the manual to illustrate various points about elicitation and rating. The language level of the text was simplified as much as was possible without sacrificing the precision of the text. The FLTB felt strongly that this simplification would help non-native English speakers to comprehend the concepts outlined in the manual more easily. Because the English tester trainees would be novices, an additional introductory section on the ILR levels, distinct from their use in elicitation and rating, was added to familiarize the testers with the levels. Additional material was added to each chapter to help the testers learn the material, including focus questions at the beginning of each chapter and review quizzes at the end of each chapter. Additional tester resource materials were added as manual appendixes, including an expanded section on examinee instructions, an updated elicitation aid, revised rating forms, and level summary pages to allow testers to review levels at a glance. The *Tester Manual* includes revisions based on feedback from the English tester training workshop.

Instructions

Based on examinee feedback during the test development process and research that indicates that examinees feel less nervous in tests where they clearly know what to expect, the FLTB created and refined throughout the Spanish and English studies a set of written examinee instructions to be read before the test as well as a set of introductory instructions to the Situation and IGT portions of the SPT. Written instructions provide three benefits: they standardize the information received by examinees; they free testers to concentrate on the upcoming task (rather than on a list of points to cover in the introduction of the task); and they avoid providing the examinee with key vocabulary. The finalized text of these instructions is included in this report as appendix A.

Additional Materials

Videotaped test segments were identified for use during training to supplement live practice tests. Between the Spanish and English studies approximately eight SPTs were videotaped. (The numerous English videos created during the English tester training workshops and the data collection phase of the English pilot study will be an invaluable resource for all future training.)

Section 5. Pilot Study Research Design

The purpose of the validation studies called for under the ULTP is to assess the reliability of the SPT as a measure of speaking proficiency. The studies were designed to evaluate the new test procedures and rating mechanisms.

The main goals for the Spanish study were the following:

- To establish a baseline for the SPT development process.
- To determine the effectiveness of the training procedures and materials initially developed.
- To establish the effectiveness of retraining for the new test.
- To gather statistical evidence of test reliability and cross-agency agreement in assigned scores.

As noted above, due to resource constraints, the original ULTP timeline was modified to include the planning and implementation of a validation study using non-native English speakers as examinees. The main goals for the English study were slightly different from those of the Spanish study and included the following:

- To determine the effectiveness of the training procedures and materials developed to date for the training of novice testers.
- To establish the effectiveness and comparability of tester training at separate sites, closely resembling future operational conditions, when SPT training will be conducted at multiple locations.
- To gather statistical evidence of test reliability and cross-agency agreement in assigned scores of the SPT in its revised form.
- To obtain videos of tests in English to be used in the training of testers in all languages.

This section describes the validation study designs and data collection procedures.

Testers

The ULTP called for FLTB member agencies that regularly perform speaking tests (CIA, DLI, FBI, and FSI) to provide personnel to participate as testers in the pilot validation studies. CIA, DLI, and FSI provided language instructors from their respective language schools, and FBI provided linguists from FBI Headquarters and Field Offices.

Spanish

From November 1994 through February 1995, 16 experienced testers, four from each of the participating agencies, came to CALL to participate in the Spanish pilot study. These testers were all previously trained in the test format currently in use at their respective agencies, and all of the testers had experience with the ILR scale and with administering tests.

The Spanish tester training workshop was held in October 1994 at the National Foreign Affairs Training Center. At this workshop, the 16 testers participating in the study were retrained in the SPT format. FLTB members and experienced tester trainers from the various agencies attended the training and took part in presenting the material. The training workshop consisted of a two-week classroom experience followed by two weeks of practice testing. During the classroom phase, testers were exposed to the principles of the new test, watched sample videos, and performed a few sample tests with tester trainers. During the practice testing, tester trainees administered tests with other trainees. Near the end of the tester training workshop, the testers were divided into eight same-agency testing pairs and each pair was assigned to one of two teams. The pairs were formed based on the expert judgment of the FLTB members and Spanish tester trainers to promote harmony and maximize reliability within the pairs. These team and pair assignments remained constant throughout the data collection phase, so testers administered SPTs with the same partner for all of the pilot testing. Results of tests administered by these pairs were compared in terms of level of agreement, and this level of agreement will be used as the basis for measuring the reliability of the SPT. The Spanish study results include some tests where one tester from the other team for the same agency served as a substitute when the usual tester was unable to test. The pattern of substitution was carefully monitored so that each of the two possible stand-ins substituted in an equal number of tests.

Because this was the first time the new SPT materials had been used, FLTB members and tester trainers spent time during the workshop requesting oral feedback from the testers about how well the training and materials were working. This feedback, as well as feedback from a number of tester meetings held early in the data collection phase, was used to improve the materials for use in the English and future studies.

English

From April 1995 through June 1995, eight government employees with no prior training as language testers came to CALL to participate in the English study. Most of these testers were drawn from the ranks of teachers and linguists from two FLTB agencies. In addition, two testers were assigned to the pilot study from other organizations, including one language instructor from the Defense Language Institute English Language Center and one member of the Department of Justice's Executive Office of Immigration Review.

Because the English study was designed to provide information on the usefulness of the tester training materials and test procedures for novice testers, it was not necessary to assign them to same-agency pairs. This type of assignment had been required in the Spanish study due to the Spanish testers' previous home-agency-based training and experience. The English pairs were formed based on the expert judgment of the tester trainers to promote harmony and maximize reliability within the pairs and remained constant throughout the data collection phase. Therefore, no interagency effect was analyzed for the English study, although inter-pair comparisons are reported as a measure of SPT reliability achieved during the English study.

The English tester training workshops were carried out in April and May 1995 using the new, revised SPT procedures and training materials. These novice testers were trained at two separate sites by interagency teams of experienced tester trainers. One workshop was held at CALL, and the other was held at the Defense Language Institute Foreign Language Center in Monterey, California. At these workshops, the testers were trained in the new SPT format, elicitation procedures, and rating procedures. Three experienced tester trainers presented each workshop. Each training workshop consisted of a two-week classroom experience followed by two weeks of practice testing, which was similar to the format used in the Spanish study. However, the trainers used the same syllabus in each of the two workshops, which had recently been revised and in which the presentation of the material had been reordered and refined. During the classroom phase, as in the Spanish study, the testers were exposed to the principles of the new test, watched sample videos, and performed a few sample tests with tester trainers. During the practice testing, tester trainees administered tests with other trainees. Since the trainees were novices, the trainers gave extensive feedback after each practice test. Near the end of the tester training workshop, the testers were divided into four testing pairs. These pair assignments remained constant throughout the length of the data collection phase, so testers administered SPTs with the same partner for all of the pilot testing. In cases where an English tester was unable to test on a given day, the tests were rescheduled, since there were no fully-trained substitute testers available.

As this was the first time the new SPT materials had been used on novice testers, FLTB members and tester trainers requested oral feedback from the testers about how well the materials were working during the training workshop. This feedback was used by the FLTB in shaping revisions of the materials for future workshops.

Subjects

The validation study design called for the administration of SPTs to volunteer examinees drawn from a pool of government employees similar to those on whom the test will be used after implementation. For this reason, recruitment of examinees from within the FLTB member agencies was emphasized. There were two goals for this recruitment approach: first, to identify government employees with language proficiency across all the levels of the ILR scale, and, second, to ensure that the validation study results would be applicable to the population that would take the test after it was implemented.

Examinees took the SPT four times, each time with a different pair of testers. Each test ran from 30 to 70 minutes. Examinees were scheduled to take their tests in two testing sessions lasting approximately three hours each—usually on different days—with each session including two SPTs and time to fill out examinee questionnaires about each test. The study design required that each examinee complete one session in the morning and one in the afternoon to counterbalance possible effects due to the time of day at which the test was administered. The study design was also carefully counterbalanced for order of testing to control for practice effect and also for examinee and tester fatigue. The testing schedule was designed so that examinees were tested in a set order by each pair. This schedule was put into place to avoid skewing results due to test order effect.

Spanish

Between November 1994 and February 1995, 138 examinees participated in the Spanish data collection phase. However, due to scheduling problems, a few examinees were unable to complete all of the required four tests. Examinee files with incomplete data were removed from the data analyzed for this study, leaving a total of 125 examinees who took 500 tests.

The ULTP called for the participation of 20 examinees at each of the six ILR levels examined (levels 1+ through 4) for the Spanish study. It was also agreed that examinee test results falling outside this range (either levels 0 through 1 or levels 4+ through 5) would be analyzed to ensure that the new SPT format would be valid for the entire ILR scale. Everyone who participated in the Spanish pilot study was a volunteer. Each of the FLTB member agencies agreed to provide as many examinees as possible from within its pool of Spanish-speaking personnel. In addition, to ensure that the proper number of examinees participated in the study, CALL also sought potential volunteers from equivalent populations at universities and other government agencies in the greater Washington, DC, metropolitan area. The scheduling of examinees was carried out by CALL testing staff. Announcements were made by FLTB members in their respective agencies and at ILR plenary sessions. A request for volunteers was posted on a number of electronic mailing lists that specialize in linguistics and foreign language teaching. Letters were sent to the entire ILR membership and to other organizations known to CALL.

Besides those provided by the six FLTB member agencies (CIA, DIA, DLI, FBI, FSI, and NSA), volunteers also participated from USIA, USAID, State Department, CIA Headquarters, FBIS, the Air Force Frequency Management Agency, the FBI Washington Metropolitan Field Office and FBI Headquarters translation staff, and the Department of Education.

English

Between May and June 1995, 86 examinees participated in the English data collection phase. However, due to scheduling problems, a few examinees were unable to complete all of the required four tests. In addition, as the study drew to a close, CALL canceled some examinees who had tested at level 2 and below because the study's goal for participation at these levels (20 examinees per level) had been met. Examinee files with incomplete data were removed from the data analyzed for this study, leaving a total of 75 examinees who took a total of 300 tests.

Although the English study was not originally included in the ULTP, the FLTB agreed to structure it similarly to the Spanish study. The FLTB's plan for the English study called for the participation of 20 examinees at each of the six ILR levels examined in the study (levels 1+ through 4). It was also agreed that examinee test results falling outside this range (either levels 0 through 1 or levels 4+ through 5) would be analyzed to ensure that the new SPT format is valid across the entire ILR scale. Government employees who participated in the English pilot study were volunteers, but non-government participants were paid $50 for participation. Each FLTB member agency agreed to provide as many examinees as possible from within its pool of non-native English-speaking personnel. In addition, to ensure that the proper number of examinees participated in the study, CALL also sought a full complement of volunteers from equivalent populations at universities, ESL institutes, and other government agencies in the greater Washington, DC, metropolitan area. The scheduling of examinees was carried out by CALL testing staff. Announcements were made by FLTB members in their respective agencies and at ILR plenary sessions. A request for volunteers was posted on a number of electronic mailing lists that specialize in linguistics and foreign language teaching. Letters were sent to the ILR membership, to all language schools in the area offering ESL classes, and to other organizations known to CALL.

Besides those provided by the six FLTB member agencies (CIA, DIA, DLI, FBI, FSI, and NSA), examinees also participated from George Mason University, US Department of Agriculture Graduate School, Department of Education, and a number of local private language teaching institutions, including Berlitz, Diplomatic Language Services, International Center for Language Studies, Inlingua, Lado International College, and the Language Exchange. Additional examinees were referred by CALL staff, FLTB members, and trainers familiar with the study.

Instructions

The FLTB members developed a set of test instructions for the examinee, consisting of an information sheet to be read by the examinee before the test as well as a script to be read aloud by the testers to the examinee at the start of the test. The instruction sheet contains the following information about the SPT: (a) format, (b) timing, (c) purpose, (d) rating criteria, (e) content, (f) outline of test activities, and (g) hints on doing well. The tester script contains summary questions and statements on the following test elements: (a) whether the examinee has read the information sheet, (b) whether the examinee has any questions about it, (c) purpose, (d) timing, (e) the right of the examinee to refuse a topic, and (f) an invitation to the examinee to take an active role in the test. These instructions are provided in English. Appendix A contains the latest version of the information sheet and the tester script.

When the testers introduced the Situation, they provided additional instructions about the Situation orally and generally asked the examinee to read a card describing the Situation. To introduce the IGT, testers handed the examinee a card with instructions for the activity and then introduced the topic orally. Because testers would have an idea of the examinee's level by the time the Situation or IGT was introduced, they were asked to give these later activity-specific instructions in English to examinees with a proficiency under level 3. Higher level examinees usually received these instructions in the test language. The IGT instructions are also included in appendix A.

Examinee Questionnaires

Two questionnaires were designed for use in the validation studies, a pretest questionnaire and a post-test questionnaire. The purpose of the pretest questionnaire was to collect basic information on the examinee's background, study and use of the test language and other foreign languages, and previous proficiency testing experiences. These background variables were considered potentially relevant to the test results. The purpose of the post-test questionnaires was to gather examinee opinions about the test. Examinees were asked what they liked and disliked about the test, whether or not they were sufficiently challenged, and whether they thought the speech sample they produced was representative of their true abilities. During the Spanish pilot, each examinee filled out one questionnaire at the end of each testing session—after the second test and after the fourth test. The examinees were asked to respond to the questions in relation to the two tests they had just completed, and, if needed, to use the space available to write comments about one specific test.

For the English study, the post-test questionnaire was further revised to make it test-specific. Examinees filled out a questionnaire after each test and were asked to comment only on the test just taken. After the final test, the examinees also filled out a summary questionnaire, in which they ranked their tests from easiest to hardest and their performance from best to worst during the four tests. Space was also provided for summary comments.

These questionnaires allowed the FLTB to collect feedback on each examinee's reaction to the new test. Once the questionnaires for the entire study had been collected and reviewed, examinee feedback was used to improve the forms. The questionnaires were revised to shorten the time it took examinees to fill out the forms and to make it easier to quantify data. Some of the questions were reworded, and some questions were replaced. A copy of the latest version of the pretest questionnaire is included in appendix B, and a copy of the post-test questionnaires are included in appendix C.

Tester Questionnaires

Near the end of each validation study, each tester was provided with an extensive questionnaire to fill out about his or her experiences in the study. The FLTB asked the testers to provide as much detail as possible about their experiences during the study, their opinions about the new test format and materials, and other aspects of speaking testing. During the data collection, testers in both the Spanish and the English studies also participated in periodic tester meetings with members of the CALL testing staff and the FLTB to discuss aspects of the study. In addition, the Spanish testers participated in a video-teleconference that reunited all of the testers to debrief the FLTB on their experiences and to answer questions from the FLTB members.

Data Collection Procedures

A strict examinee testing schedule was established, with each testing pair administering four tests per day. During the Spanish study, each testing pair administered tests on two days each week so that both teams of testers could participate; during the English study, the testing pairs administered tests four days per week.

Spanish

The testing pairs administered and scored a total of 500 SPTs for 125 Spanish speakers of varying proficiency levels; videotapes of a randomly-selected sample of tests were also re-rated by the corresponding agency pair on the other team.

Because of constraints on the length of time the testers could be detailed to participate as Spanish testers, it was necessary for each testing pair to perform four tests per testing day. Each of the four participating agencies detailed four testers, or two testing pairs, for the Spanish study. One pair from each agency was assigned to one of two teams. Each team of four pairs tested the same group of examinees on the same days. The team 1 pairs tested Monday and Wednesday of each week, while the team 2 pairs administered their tests on Tuesday and Thursday. The testing was arranged in this way, with one day on and one day off, to reduce tester fatigue and prevent burnout. When testers were not administering tests, they addressed testing issues in group meetings facilitated by a trainer, met with CALL staff and FLTB members in tester meetings, developed new testing materials, or met in small groups on their own. The pilot testing ran for 12 weeks, beginning in November 1994, with a two-week break in mid-December. Testing began again in early January and was completed in mid-February 1995.

During the Spanish study, testers were given a number of instructions about the administration of SPTs by the FLTB to ensure consistency in the testing: (1) testers were asked to place the activities during the test in the following set order: Conversation, Situation, IGT, (2) testers were asked to use the approved list of Situations when presenting Situations during the test using a card. Situations introduced without a card did not need to be on the list, (3) testers were asked to provide instructions in English except in the case of high-level native speakers—when they could either present the instructions in Spanish or in English, and (4) both testers were asked to remain in the room during the IGT, except in a few tests of high-level native Spanish speakers.

English

After being trained in separate workshops, the new testers were brought together at CALL to begin the pilot test data collection. The eight testers were divided into four pairs to carry out the testing in May and June 1995. The testing procedures were essentially equal to those of the Spanish study, with a few exceptions.

Effect of Tester's Presence/Absence During the IGT. In the case of English testing, since the test language also had to serve as the common language of the examinees and testers, some adjustments were made to the procedures to make the use of English more realistic and to provide clearer instructions to the examinee. The written instructions were simplified considerably so that examinees with low-level reading ability in English would find them easier to understand. In the IGT, one tester left the room during the time the examinee was interviewing the other tester. It was felt that since the examinee had to elicit information in English and report to the other tester using English (in other words, since no "interpreting" was necessary), it would be unnatural to have both native-English-speaking testers in the room throughout the procedure.

Situation/IGT Placement. During the English pilot, some testing pairs chose to place the IGT before the Situation during their tests. Those who did this felt that examinees tended to misunderstand the purpose of the IGT by treating it as another Situation to be acted out. Introducing the Situation after the IGT helped to reduce any confusion the examinees may have felt about what was required of them during the IGT procedure.

Section 6. Rating Reliability Results

The pilot validation studies were designed to answer important questions about the new test format. This section will contain the results on the use of the ILR Scale within the Spanish and English studies as well as SPT reliability results from a number of statistical analyses. For these analyses, the ratings were coded with base levels at 00, 10, 20, 30, 40, and 50 and plus levels set at .8, so that plus levels were 08, 18, 28, 38, and 48. This coding was done in keeping with discussions by the FLTB on the historical precedent, the nature of the scale, and the characteristics of plus levels.

Use of the ILR Scale

The first area the Spanish and English pilot studies were designed to examine was how the testing pairs used the ILR scale during each study. As was described above, the Spanish testers were assigned to agency-specific pairs. The English testers were novices with no previous background in language testing. They were assigned to mixed-agency pairs. For this and the remaining sections of the report, the Spanish study results will be reported in terms of interagency comparisons, while the English study results will be reported in terms of inter-pair comparisons.

Frequency Charts: Spanish and English Pilot Studies

Descriptive analyses were run to create frequency tables showing the distribution of final scores for all of the tests administered during the Spanish study. Charts were created for the study overall, for teams 1 and 2, and for the first and second halves of the study. Similar charts were created for the English study, illustrating the frequency distributions for the study overall, for the first and second halves of the study, and for each of the testing pairs. The charts of these distributions are included at the end of this document as appendix D.

Normality Results: Spanish and English Pilot Studies

Five additional types of data were provided to evaluate the normality of the charted frequency distributions, that is, whether the data distribution fell into a pattern that would fit under a bell-shaped curve. Tables containing this data have been prepared and placed in appendix D under each chart. These tables report (1) the median score assigned as well as (2) the interquartile range for each chart. These data indicate the extent to which the final ratings assigned during the studies were spread out across the ILR levels. The tables indicate that the interquartile range (IQR)—the difference between the score assigned at the 75th percentile and that assigned at the 25th percentile—is greater for the English study results than for the Spanish study results. This difference may be due to the use of novice testers in the English study. In addition, the numerical values of (3) skewedness and (4) kurtosis were also reported for each chart. Last, each table contains the results of (5) a K-S Lilliefors statistic, which tests the distribution of the data in each chart against a normal distribution. A significance (or p) value of less than .05 means the distribution is non-normal (Norusis, 1994). The Lilliefors statistic seemed to be hypersensitive to non-normality, in that it found all of the final rating distributions for both studies to be non-normal. However, taking into account the results of the other measures, English distributions tended to be non-normal while almost all of the Spanish distributions were normal.

Note on Measures Used in Reliability Section

This section will provide a brief description of and assumptions for the analyses used, and it will outline the reliability research questions selected for examination in this report. It will also provide summary tables containing the results of each analysis for each question, and, last, it will provide a brief interpretation of those results. Reliability was measured and reported in the following section using percent level of agreement as well as a number of non-parametric statistical measures.

Percent Level of Agreement

One question that arises naturally in any situation where language test performance is rated is how often raters should agree for a rating to be considered reliable. This section will describe the method selected for determining rater reliability, which has come to be known in the literature as **percent-agreement**.[1] This method computes rater agreement as a percentage. For example, if two raters judge 100 students and agree on the precise rating for 50 students (e.g., rater A awards student X a score of 1, and so does rater B), then the raters would be able to report an achieved 50% level of agreement. If the two raters agree on 75 students, then their percent level of agreement would be 75%. Appendixes G and H contain cross-tabulation charts for the Spanish and English studies. The shaded cells on these charts contain the number of examinees for whom the raters agreed exactly on the various ILR rating levels.

---

[1] Percent level of agreement between agencies and testing pairs will be reported as a measure of reliability in this report. The decision to use percent agreement rather than an alternative, called Cohen's kappa, was made for a number of reasons. First, percent-agreement is a more easily interpreted statistic than kappa (see below for a description of this statistic). Second, although some researchers in the area of rater agreement have over the years utilized the kappa statistic, this statistic has not achieved widespread use and common interpretation among statisticians, and it still is not regularly reported in foreign language test reports. Research reports can still be found where simple percent-agreement is reported in lieu of the kappa statistic (e.g., Cole, et al., 1991; Nugent and Loabs, 1978; Schroeder, 1973), although kappa has been reported in some testing research (Clapham, 1994; Thompson, 1995).

In his seminal article that proposed this statistical approach, Cohen labeled the reporting of reliability in terms of percent-agreement as "primitive" (Cohen, 1960: 38) and proposed a statistic called kappa instead. Kappa "is simply the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement *after* chance agreement is removed from consideration" (Cohen, 1960: 40, emphasis in original). Put another way, kappa is a statistical test for agreement along the diagonals of a rater-by-rater agreement frequency table. (The diagonal agreements for the pilot study data are represented as shaded cells in the cross-tabulation charts included in appendixes G and H.) Cohen's kappa statistic, unlike the non-parametric chi-squared statistic, is an inferential statistical test of whether the diagonal agreements in a table are greater than would be expected by chance. If those diagonal frequencies are exactly as expected by chance, then kappa would be zero. If those frequencies are greater than expected by chance, then kappa is above zero. Kappa can be interpreted much as a correlation coefficient, in that, if the raters agree perfectly, then kappa achieves its maximum possible value: +1.00.

A variant of kappa, called weighted kappa, has also been reported in some research. "Kappa is useful when all disagreements may be considered equally serious, and weighted kappa is useful when the relative seriousness of the different kinds of disagreement can be specified" (Fleiss, 1971: 378). Since this weighted kappa analysis could be set up to give more weight to ratings more than one rating level apart (Everitt, 1968) and was used in this way in Thompson (1995) and Clapham (1994), this statistic was also examined for potential use in this report to weight differences of more than one ILR level.

Percent-agreement cannot be tested statistically for probability, as is the case for other statistical analyses reported to address other research questions in this report. It is therefore important to determine a benchmark for selecting an appropriate level of percent agreement necessary for the SPT to be considered reliable. It is proposed for purposes of this report that a general lower-bound acceptable value for percent level of agreement should be set at 70% among SPT raters.[2] This bound is quite conservative and, if reached, should allow the SPT to be considered fairly reliable. Interagency comparisons where each agency's rating was compared individually to the ratings of every other agency for each examinee resulted in within-level percent-agreement that ranged from 50% to 63% for the Spanish study and from 51% to 64% for the English study. In 1986, a study was conducted by the Center for Applied Linguistics (CAL) with participation by CIA, DLI, and FSI. Similar comparisons run on French and German results from the CAL study reflected within-level percent-agreements of 35% to 56% and 34% to 46%, respectively. Although the SPT 1-to-1 comparison results do not meet the 70% cutoff, they do represent about a 10% improvement over the results of the 1986 CAL study. Looking exclusively at the SPT studies, the Spanish results met this 70% level of agreement (72%) when within-level agreements among three of four agencies were considered, and the English study results reached 64% for within-level agreement for three of four agencies.

Non-Parametric Statistical Analyses

The data from the Spanish and English pilot studies were also analyzed through the use of three non-parametric statistical measures. Information on these analyses was drawn from Hatch and Lazaraton (1991) and the various Norusis (1994) references included in the bibliography of this report. Non-parametric analyses are appropriate for ranked data. As was discussed in the test procedures section, the ILR scale is made up of six ranges of language proficiency, beginning with a rather narrow range at level 0 and increasing substantially for each successively higher level. For example, the range covered by level 3 will be much larger than the range covered by level 2. Each higher level also includes the language functions, tasks, and characteristics described at all lower levels, so that, for example, level 4 subsumes the skills and abilities described for levels 0 through 3 in addition to the new set of requirements for level 4. It has been argued that ILR ratings, with an expanding range at each successively higher level, function more as points along an ordinal (or ranked) scale than as points along an interval scale, which assumes equal distance between intervals (Hart-Gonzalez, 1993). One further characteristic of the pilot study data that affects the choice of statistical tests is that the distributions of the ratings in one case for the Spanish pilot data and in all cases for the English pilot data were found to be non-normal. These normality test results provide further support for the use of non-parametric statistical analyses, given that parametric tests generally assume that the data are distributed in a normal or bell-shaped curve. For these reasons, only non-parametric statistical analyses will be reported in this report.

---

[2] As noted previously, the kappa statistic will not be reported; however, published comparisons have been made between percent-agreement and kappa in an effort to judge what percent-agreement value would generate significantly non-zero kappa coefficients. Review of that literature reveals some interesting conclusions. First, a number of studies (Cole, et al., 1991; Nugent and Loabs, 1978; Schroeder, 1973), which report percent-agreement statistics only, provide values that range from 67% to 100% and appear to treat them as acceptable. Second, more mathematically tractable approaches have directly examined the relationship between kappa and percent agreement. Some researchers found that kappa does not rise above zero until percent agreement is about .70 in situations with two raters (Umesh, et. al. 1989: 844; in this regard, see also Mason, 1992: 352). This recalls a comment by Cohen in his original paper on kappa: ". . . it is generally of as little value to test kappa for significance as it is for any other reliability coefficient—to know merely that kappa is beyond chance is trivial since one usually expects much more than this in the way of reliability in psychological measurement. It may, however, serve as a minimum demand in some applications" (Cohen, 1960: 44). Another study reports high and strong kappas when percent-agreement reaches at least 66.1% (Kaplan and Johnson, 1992: 15). The kappas reported there would all easily achieve inferential significance, and, if interpreted as correlation coefficients, they would indicate a strong positive relationship.

The level of significance (α) selected for this project is .05 in accordance with customary statistical procedures (Hatch and Lazaraton, 1991). This means that the odds of the results being due to chance are 5 in 100. In the tables that follow, results for which the probability values meet this level of significance will be marked with a single asterisk (*). Results for which the probability values reach an even higher level of significance, such as .01 (1 in 100), will be marked with double asterisks (**).

The following tests were selected as most appropriate for analysis of the data. A description and justification for the use of each of these tests appears below:

- Non-Parametric Correlation: Kendall's Tau-b
- Non-Parametric Exact-Test Chi-Squares
- Non-Parametric Analyses of Variance:
    - —Friedman Chi-Square of Ranks Test
    - —Wilcoxon Matched-Pair Signed-Ranks Test
    - —Sign Test

Many of the tests reported on compare two variables at a time, so the display of the results will be presented in the form of a matrix, with individual cells on the table corresponding to the results of the comparison of the agencies located on the row and column for that cell. Each table will also contain information on the specific analyses run.

Sample Test Results Format

	FSI	**FBI**	**DLI**
CIA	Results of analysis comparing CIA & FSI	Results of analysis comparing CIA & FBI	Results of analysis comparing CIA & DLI
DLI	Results of analysis comparing DLI & FSI	Results of analysis comparing DLI & FBI	
FBI	Results of analysis comparing FBI & FSI		

Explanation of the table, including the name of the statistical analysis for which results are being reported, a description of the groups being compared, and an explanation of headings used in the table.

The sections below will provide a brief description of each test used as well as a brief explanation of why it was selected.

Kendall's Tau-b Correlation. The most commonly used non-parametric correlation formula in applied linguistics research is the Spearman Rank–order correlation. However, it is generally considered that this formula should not be used on data that contain a number of tie rankings. Instead, Kendall's tau formula b is recommended in these cases (Hatch and Lazaraton, 1991). The examinees' performances were assigned to 1 of 11 ranked categories on the ILR scale in both the Spanish and English studies. If Spearman Rank–order analyses were used, the computer would treat all of the examinees assigned to a given level as ties. Kendall's tau-b includes an adjustment for ties in the data set (Norusis, 1994), so it is the most appropriate formula for the SPT pilot study data and will be the type of correlation used to calculate reliability coefficients throughout this report.

Non-Parametric Chi-Square Using Exact Test. The non-parametric Pearson chi-squares test analyzes the frequency distributions of two variables to determine whether the differences between the two distributions are statistically significant. For example, the distribution of ratings assigned by CIA can be compared to those assigned by DLI to see if there were statistically significant differences between them. Since one of the major goals of the ULTP is to remove or reduce rating differences across the agencies, it is important to identify areas in which differences occur. In these studies, some of the sample sizes assigned to the matrices being analyzed are small, so that the expected cell frequencies used in the calculations often fall below 5, the minimum number usually needed for verifiable results. A special formula for computing these chi-squares analyses was used to overcome the problem of lower cell frequencies: the Monte Carlo algorithm in the SPSS Exact Tests module.

Friedman Chi-Square of Ranks Test. The Friedman test is considered the non-parametric parallel to a repeated-measures ANOVA. The Friedman test allows for comparison of two or more groups of ranked data. For example, this test will be used to analyze whether the four agencies differed significantly in how they assigned final ratings. The test provides information as to whether there are significant differences among the groups, but it does not indicate where the differences are.

Wilcoxon Matched-Pair Signed-Ranks Test. While the Friedman test is the non-parametric repeated-measures ANOVA, the Wilcoxon test is one non-parametric parallel to the matched t-test. In cases where the Friedman test indicates differences among two or more groups, a Wilcoxon test will provide information about which groups are statistically different from one another. This test analyzes the rankings of each examinee to see whether differences in the rankings exist, assesses the direction of any change, and measures the degree of change. This test weights differences of more than one level more heavily than differences of a single level.

Sign Test. The Sign test is another non-parametric parallel to the matched t-test, but it provides information only about the existence and direction of change. The results of both the Wilcoxon and Sign tests will be provided in appendix E for the Spanish data and in appendix F for the English data.

Spanish Pilot Study

The Spanish study was set up to provide a close look at the separate agencies' performance, in that it was anticipated that the Spanish testers' previous experience with their respective agency's testing procedures might have some effect on their performance of the SPT procedures. It was believed that testers would overcome difficulties in eliciting a ratable sample by relying upon their previous training and experience, which could have an effect on the study results. The testing pair who conducted the test provided what will be referred to below as the **live** rating. Each test was videotaped and audiotaped. A random selection of videotaped tests was re-rated by the same-agency pair on the other team. The ratings of videotapes, referred to below as **taped** ratings, took place in conditions as similar to the live ratings as possible, in that testers were asked to view each test in its entirety and provide a rating in one uninterrupted session, following the same rating procedures that they would use to rate their own live tests. The results from the taped ratings were analyzed to provide additional information about test reliability within agencies.

The following questions are addressed in the sections below:

- **Interagency reliability**:
  How well did the agencies agree on their final ratings for each examinee?
- **Intra-agency reliability**:
  How well did the first, live ratings from each testing pair from each agency agree with the second, taped ratings assigned by the other pair from their own agency?
- **Inter-rater reliability**:
  How well did the testers in each pair agree with one another on each test?
- **Effects on reliability caused by test order and time of administration**:
  Was there an effect on ratings caused by test order?
  Was there an effect on ratings caused by the time of day when the test was administered?

The sections below will address each of these research questions in turn. The results will be reported for analyses that have been conducted on various subsets of the data. For live ratings, the results of the overall study will be reported, taking into account all of the data from the study. Results will also be reported for the separate groups of examinees tested by each of the teams (teams 1 and 2) and for the separate groups tested during the first and second halves of the study (phases 1 and 2). For taped ratings, only overall results will be reported, since the number of examinees selected for taped ratings is too small to subdivide further.

Interagency Reliability

This section will report on the results of analyses conducted to assess the amount of and patterns of interagency agreement and disagreement found among the final negotiated ratings for the Spanish tests. One of the most important benefits and perhaps the main goal of this effort of creating and implementing a common speaking proficiency test is to ensure that a single examinee taking the new test will receive the same rating—no matter which agency administers the test. For this reason, it is expected that when the SPT is fully implemented, with joint training on a single set of test procedures, no significant differences would be found among the ratings by the different groups. Cross-tabulation charts for the distribution of final ratings are included in this report as appendix G. The following sections provide data on how closely the Spanish pilot comes to this ideal.

Agency Rating Analyses. This section will report on the level of agreement among the agencies on each examinee. The results of the Spanish pilot study from these rating analyses are included in the following tables.

Agency Rating Analyses-Exact Matches: Spanish Pilot Study

	N	Exact Matches (4)	Exact Matches (3)	Exact Matches (none)[3]
Overall	125	12 %	30 %	5 %
Team 1	63	11 %	29 %	6 %
Team 2	62	13 %	32 %	2 %
Phase 1	57	11 %	29 %	2 %
Phase 2	68	13 %	32 %	6 %

Exact matches (4) *includes the percentage of examinees for whom all agencies assigned exactly the same score.* ***Exact matches (3)*** *includes the percentage of examinees for whom at least three agencies assigned exactly the same score (including the percentage for whom all four agencies agreed exactly).* ***Exact matches (none)*** *indicates the percentage of examinees for whom all agencies assigned a different final score. The* ***overall*** *results take into account all tests administered during the Spanish study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The percent-level of exact agreement among four testing pairs remained relatively constant throughout the study, while the exact agreement among three testing pairs seemed slightly higher for team 2. It also seems that there were slightly more exact matches during the second half of the study than during the first half. The percentage of examinees for whom no agencies agreed also fluctuated across teams and phases. These differences may reflect the nature of the examinees who participated in the two halves of the study rather than being due only to tester behavior.

---

[3] Note that the percentages reflect a total of five examinees for whom none of the pairs agreed. The relative differences in the percentages are a function of the total number of examinees (N) included in the analyses.

The following table reports on the within-level percent-agreement among four or three agencies

Agency Rating Analyses-Within-Level Matches: Spanish Pilot Study

	N	Within-Level Matches (4)	Within-Level Matches (3)
Overall	125	30 %	72 %
Team 1	63	29 %	70 %
Team 2	62	32 %	74 %
Phase 1	57	39 %	75 %
Phase 2	68	24 %	71 %

Within-level matches (4) *includes the percentage of examinees for whom all agencies assigned exactly the same score plus those where all of the ratings fell within a given level, that is, where all four agencies assigned either a given ILR base level or its respective plus level (e.g., all 2 or 2+ ratings).* ***Within-level matches (3)*** *includes the percentage of examinees for whom at least three agencies assigned scores within the same level (plus the percentage for whom all four agreed exactly and within-level, and when three agreed exactly). The* ***overall*** *results take into account all tests administered during the Spanish study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The percentage of within-level four-agency agreement was higher during the first half of the study than during the second half of the study. The percentage of three-agency agreement was also highest during the first half of the study. These differences may reflect the nature of the examinees who participated in the two halves of the study rather than being due only to tester behavior.

For the following tables, average percent level of exact agreement was calculated for each agency by comparing that agency's rating for each examinee with those assigned by each of the other participating agencies.

Agency Rating Analyses
Percent Level of Agreement by Agency (Exact Matches):
Spanish Pilot Study

	CIA	DLI	FBI	FSI	Average
Overall	36%	38%	36%	38%	37%
Team 1	37%	38%	34%	37%	37%
Team 2	34%	38%	39%	39%	37%
Phase 1	40%	39%	38%	41%	40%
Phase 2	33%	38%	35%	37%	36%

Ratings assigned to a given examinee by each agency were compared to those assigned by each of the other agencies individually, e.g., CIA's percent level of agreement was calculated by averaging CIA's percentage of agreement with DLI, with FBI, and with FSI. The ***average*** *column reports the average for the overall study.* ***Exact matches*** *includes the percentage of examinees for whom the two agencies assigned exactly the same score. The* ***overall*** *results take into account all tests administered during the Spanish study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

These results indicate that the percent level of agreement between the ratings by the two sets of testing pairs assigned to the teams is more or less equal. The phase analyses indicate a slightly lower level of agreement in phase 2 than in the first half of the study.

The following table reports similar comparisons of each agency to every other while reporting within-level agreements.

Agency Rating Analyses
Percent Level of Agreement by Agency (Within-Level):
Spanish Pilot Study

	CIA	DLI	FBI	FSI	Average
Overall	52%	59%	58%	59%	57%
Team 1	55%	59%	58%	57%	57%
Team 2	49%	56%	59%	61%	56%
Phase 1	60%	61%	63%	64%	62%
Phase 2	46%	58%	56%	56%	54%

Ratings assigned to a given examinee by each agency were compared to those assigned by each of the other agencies individually, e.g., CIA's percent level of agreement was calculated by averaging CIA's percentage of agreement with DLI, with FBI, and with FSI. The ***average*** *column reports the average for the overall study.* ***Within-level matches*** *includes the percentage of examinees for whom the two agencies assigned scores within the same base level (plus the percentage for whom the pairs agreed exactly). The* ***overall*** *results take into account all tests administered during the Spanish study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The results in this table show approximately the same pattern as that for exact matches, with the testing pairs in each team behaving slightly differently for each agency but within a narrow range of variance. The phase 2 results are also lower than those of phase 1.

Statistical Analysis of Live Ratings. This section will report on the interagency percent level of agreement, statistical similarities and differences found by various non-parametric analyses, and reliability coefficients for the same groups' live ratings. For additional details on the results, please see the Spanish pilot data summary tables in appendix E.

The following table reports the interagency percent level of agreement for the Spanish study.

Summary of Interagency Percent Level of Agreement: Spanish Pilot Study

	Low Percent-Agreement	High Percent-Agreement	Δ
Overall	38%	46%	8%
Team 1	32%	46%	14%
Team 2	39%	53%	14%
Phase 1	41%	50%	9%
Phase 2	30%	46%	16%

This table reports the lowest and highest percent level of agreement found among the comparisons made of pairs of agencies in the Spanish pilot study. The ***overall*** *results take into account all tests administered during the study, the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively. In an ideal world, all of the pairs would have been found to have 100% agreement.*

The percent level of agreement was slightly higher for team 2 than for team 1 although the percentages varied within about the same range. During the first half of the study, the percent level of agreement varied much less than during phase 2. The percent-agreement fell slightly during the second half of the study.

The following tables report on the results of a number of non-parametric analyses. The Pearson chi-squares analyses run to detect differences in how the ratings were distributed across the scale by the four agencies showed that there were statistically significant differences among the four groups for the study overall, both teams, and both phases of the study. When Friedman analyses were run to compare the four agencies to one another, they indicated statistically significant differences among the groups. A significant Friedman result indicates that there are differences among the groups, but does not identify where the differences can be found. Two additional tests, Wilcoxon and Sign, were run on each set of data from two agencies, comparing each agency to every other agency to determine the nature of the differences among the groups. Since the goal of the ULTP is to decrease statistically significant differences in ratings across agencies, the ideal result for the table below would be for all of the comparisons to show as "Same."

Summary of Interagency Wilcoxon/Sign Results: Spanish Pilot Study						
	FSI		**FBI**		**DLI**	
Overall Study						
CIA	Different		Different		Different	
DLI	Different		Same			
FBI	Different					
	FSI		**FBI**		**DLI**	
	Team 1	**Team 2**	**Team 1**	**Team 2**	**Team 1**	**Team 2**
CIA	Different	Different	Same	Different	Different	Different
DLI	Mixed	Different	Different	Same		
FBI	Different	Different				
	FSI		**FBI**		**DLI**	
	Phase 1	**Phase 2**	**Phase 1**	**Phase 2**	**Phase 1**	**Phase 2**
CIA	Different	Different	Same	Different	Mixed	Different
DLI	Mixed	Different	Same	Same		
FBI	Different	Different				

This table reports a summary of the results of two non-parametric analyses of variance: the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In this case, the final ratings assigned by the four agencies were compared two at a time. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs,* ***different*** *indicates that both tests found a statistically significant difference between the pairs, and* ***mixed*** *indicates that the tests returned different results. The* ***overall*** *results take into account all tests administered during the Spanish study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results report on tests administered in the first and second halves of the study, respectively.*

As can be seen from the table above, the pattern of differences changed slightly depending upon the subset of the data being analyzed. For the overall study, all of the agencies were found to be significantly different from one another except DLI/FBI. During the first phase of the study, the groups seemed to be closer together, in that DLI/FBI and CIA/FBI showed no statistical difference on both tests, and the CIA/DLI and DLI/FSI comparisons showed mixed results (where one of the tests indicated a significant difference while the other test did not). These pattern differences indicate that some tester drift occurred over the four months of data collection and that this drift seems to have affected most if not all of the teams.

The results of the Kendall's tau-b correlations also reflect some evidence of drift, in that the range of the correlations for all agencies is largest for phase 2.

Summary of Interagency Correlation Results: Spanish Pilot Study			
Data subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	.737	.799	.0620
Team 1	.750	.801	.0510
Team 2	.716	.805	.0890
Phase 1	.747	.809	.0620
Phase 2	.701	.798	.0970

This table reports the lowest and highest interagency Kendall's tau-b correlation coefficients for the final negotiated ratings assigned by the four agencies when they were compared two at a time. The column labeled ***Δ*** *reports the difference between the two correlation columns. The* ***overall*** *results take into account all tests administered during the Spanish study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

Another pattern discernible in the data is related to the interquartile range (IQR) around the various agency medians when the data is grouped by team and phase.

Summary of Interagency Median and Interquartile Range Results: Spanish Pilot Study				
Data Subset	**Low Median**	**High Median**	**Low Interquartile Range (IQR)**	**High Interquartile Range (IQR)**
Overall	2	2+	10.0	18.0
Team 1	2	2+	10.0	18.0
Team 2	2	2+	10.0	18.0
Phase 1	2	2+	10.0	12.0
Phase 2	2+	3	10.0	20.0

This table reports the lowest and highest median and interquartile range calculated on the Spanish pilot study data for the four agencies. The median is a measure of central tendency, and the interquartile range is a measure of the dispersion of the final ratings across the ILR scale. The ***overall*** *results take into account all tests administered during the study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The differences in medians and IQRs for the study indicate that ratings generally varied from a plus point to a full level up or down. There are also differences in the above measures for the phases; during the first half of the study, the IQR is narrower, while during the second half of the study, the IQR widened. The median was also higher during phase 2, which may be due to the characteristics of the examinees tested during that phase rather than to tester behavior.

Statistical Analysis of Taped Ratings. Statistical analyses were run on just the four taped ratings assigned to each examinee to identify the level of interagency reliability for these ratings.

Summary of Interagency Percent Level of Agreement—Taped Ratings Only: Spanish Pilot Study

	FSI	**FBI**	**DLI**
CIA	30%	33%	42%
DLI	38%	48%	
FBI	29%		

In the case of the taped ratings, a subset of live tests administered by one testing pair were re-rated via videotape by the other pair from the same agency. This table reports the interagency percent level of agreement for the taped comparisons only. In an ideal world, all of the pairs would have been found to have 100% agreement.

In terms of these calculations, it seems that FBI and FSI were least likely to agree while DLI and FBI were most likely to agree.

The following paragraphs will report the results of a number of non-parametric analyses of variance. The Pearson chi-squares test indicated that the taped ratings from each agency were distributed across the scale differently. The Friedman test also indicated differences among the groups. The following table will report the results of Wilcoxon and sign tests performed to identify where the differences were.

Summary of Interagency Wilcoxon/Sign Results—Taped Ratings Only: Spanish Pilot Study

	FSI	**FBI**	**DLI**
CIA	Different	Same	Mixed
DLI	Same	Same	
FBI	Different		

This table reports a summary of the results of two non-parametric analyses of variance: the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In the case of the taped ratings, a subset of live tests administered by one testing pair in the Spanish study were re-rated via videotape by the other pair from the same agency. This table reports the results of the taped comparisons only. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs,* ***different*** *indicates that both tests found a statistically significant difference between the pairs, and* ***mixed*** *indicates that the tests returned different results. In an ideal world, all of the pairs would have been found to be the same, with no statistically significant differences.*

In the Wilcoxon and Sign test results for these ratings, CIA/FSI and FBI/FSI were found to be different and CIA/DLI showed a difference on one of the tests. The correlations comparing the agencies taped ratings to one another were spread across a slightly wider range and were slightly lower than the reliability levels for live ratings, ranging from .630 to .814. These differences may be tied to the lower number of examinees taken into account in these results (about 24) than for the live ratings (about 125). As was mentioned above, tests were run only for the study overall rather than for the subsets of the data reported for live ratings because there were not enough taped ratings performed to divide the data set further

Intra-Agency Reliability

For the Spanish pilot study, a random sample of examinees was selected from the total examinee population for taped rating. For these examinees, each test administered by the four testing pairs was re-rated by the other pair from the same agency. These taped ratings provide additional information on intra-agency agreement and reliability. Results will be reported on comparisons of agreement between the live final negotiated ratings and the taped final negotiated ratings and correlations of reliability between the individual tester ratings assigned during the taped rating process. Please see appendix G for the expanded results tables.

<table>
<tr><th colspan="4">Summary of Intra-Agency Percent Level of Agreement: Spanish Pilot Study</th></tr>
<tr><th>CIA</th><th>DLI</th><th>FBI</th><th>FSI</th></tr>
<tr><td>49%</td><td>43%</td><td>36%</td><td>44%</td></tr>
</table>

This table reports the percent level of agreement between live ratings and taped ratings for each agency in the Spanish pilot study. These percentages reflect the level of agreement between live ratings assigned by one testing pair from each agency and taped ratings assigned by the other pair. In an ideal world, all of the pairs would have been found to have 100% agreement.

The percent-agreement for second ratings tended to be spread across a narrower range than that for live ratings. They were neither higher nor lower overall than the percentages for the live ratings. This result may be due to the lower number of examinees taken into account in these results (about 24) than for the live ratings (about 125).

The Pearson chi-squares results indicated significant differences among the live and taped ratings for all agencies. The results for the CIA, DLI, and FSI comparisons were significant at the .001 level or above, while FBI's results reached the .05 significance level. A Friedman test run on the subset of examinees selected for taped rating comparing the eight test ratings (four live ratings and four taped ratings) identified statistically significant differences among the groups. Wilcoxon and Sign tests run to compare only each agency's live rating with its taped ratings (rather than on every possible combination) are reported in the table below.

<table>
<tr><th colspan="4">Summary of Intra-Agency Wilcoxon/Sign Results: Spanish Pilot Study</th></tr>
<tr><th>CIA</th><th>DLI</th><th>FBI</th><th>FSI</th></tr>
<tr><td>Different</td><td>Same</td><td>Same</td><td>Same</td></tr>
</table>

This table reports a summary of the results of two non-parametric analyses of variance: the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In the case of the taped ratings, a subset of live tests administered by one testing pair in the Spanish study were re-rated via videotape by the other pair from the same agency. This table reports the results of the comparisons of the taped ratings with their respective live ratings. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs.* ***Different*** *indicates that both tests found a statistically significant difference between the pairs. In an ideal world, all of the pairs would have been found to be the same, with no statistically significant differences.*

As shown in the above table, no statistical differences were found for DLI, FBI or FSI, while CIA showed statistically significant differences for both tests. As for reliability coefficients, the intra-agency reliability was slightly lower than that of interagency comparisons, ranging from .652 to .770.

Inter-Rater Reliability

This section will examine the level of within-pair agreement in individual final ratings.

Statistical Analysis of Live Ratings. Reliability results will be reported in terms of percent level of agreement as well as correlations for each tester's individual rating with his or her testing partner in live ratings. Please see the summary results tables in appendix G for additional detail on these results.

Summary of Inter-Rater Percent Level of Agreement: Spanish Pilot Study

	CIA	DLI	FBI	FSI	Average
Overall Study	76%	86%	76%	99%	84%
Team 1	71%	100%	71%	100%	86%
Team 2	82%	73%	81%	98%	84%
Phase 1	71%	80%	64%	100%	79%
Phase 2	81%	91%	85%	100%	89%

This table reports the percent level of agreement between live individual tester ratings within testing pairs in the Spanish pilot study. The column titled ***average*** *provides average inter-rater percent-agreement for the Spanish study overall. The* ***overall*** *results take into account all tests administered during the English study; while the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively. In an ideal world, all of the pairs would have been found to have 100% agreement.*

The percent level of agreement varied across teams without noticeable patterns. It seems that the phase 1 percentages tended to be lower than those for the overall study and for phase 2. The percentages for team 1 varied across a wider range than those for team 2; however, these differences could be due to differences in the examinee populations tested by the different teams.

The following table reports the correlation coefficients for inter-rater reliability.

Summary of Inter-Rater Correlation Results: Spanish Pilot Study			
Data Subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	.893	.998	.1050
Team 1	.864	1.000	.1360
Team 2	.916	.995	.0790
Phase 1	.841	1.000	.1590
Phase 2	.939	1.000	.0610

This table reports the lowest and highest inter-rater Kendall's tau-b correlation coefficients for the individual tester ratings assigned within testing pairs by the four agencies when they were compared two at a time. The column labeled ***Δ*** *reports the difference between the two correlation columns. The* ***overall*** *results take into account all tests administered during the Spanish study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

These results indicate that the pairs of testers assigned to team 1 tended to disagree with one another slightly more often than the team 2 testers. It is also interesting to note that testers tended to disagree more during phase 1 than phase 2. This may indicate that as they became accustomed to testing together, they tended to agree more frequently. Based on the interagency results above, though, they seemed to move closer to one another but further from the other agency pairs. It was recognized that there was the possibility of testers becoming familiarized with one another since they were assigned to static testing pairs for the entire data collection phase. Furthermore, this phenomenon may occur during operational testing in agencies where testers consistently test with one other person. Such tester drift within teams should be identified in each agency and corrected through retraining or rotation with other testing partners.

Statistical Analysis of Taped Ratings. Statistical analyses were also run on the inter-rater reliability of ratings assigned by testers during the taped ratings. The following table contains data related to the correlations for individual tester ratings on taped ratings only.

Summary of Inter-Rater Correlation Results—Taped Ratings Only: Spanish Pilot Study	
Lowest Correlation	**Highest Correlation**
918	.996

This table reports the lowest and highest Kendall tau-b correlation coefficients between individual tester ratings in the Spanish pilot study for taped ratings only.

Percent agreement was slightly higher and was limited to a narrower range than that for live ratings. Correlations were slightly higher as well. This may be due to the reduced number of examinees (24) included in these analyses compared to that used in the analyses of live ratings (125).

Effects on Reliability by Test Order/Time of Administration

This section will report on the results of analyses conducted to assess the amount and patterns of agreement and disagreement found among the final negotiated ratings for the Spanish tests, examining the tests by test order and time of administration. It is important to note that the data collection schedule was designed to counterbalance for variance due to test order and timing effects by spreading this variance across all teams; however, this data may be of interest to program managers as they arrange testing schedules. Potential sources for variance among the groups include examinee practice effect, examinee fatigue, and tester fatigue. It was expected that examinees would become better at performing the different sections of the SPT with multiple administrations and that perhaps their scores would improve slightly, but it was also believed that the act of taking two tests one after another would tire examinees and reduce their scores slightly. There was also some concern about potential effects from tester fatigue as well.

Test Order. Test order effects were analyzed by grouping the final negotiated ratings for every examinee's first test in a single group, all second tests in a different group, and so on for their third and fourth tests. This section will examine the level of agreement and differences between these groups for the overall study, teams 1 and 2, and for phases 1 and 2. It will also provide reliability coefficients for the same groups. This section will take into account live ratings only. Please see appendix E for more detail on these results.

Pearson chi-squares tests run on the distribution of ratings for first, second, third, and fourth tests showed that there were statistically significant differences among all of the groups for the overall study, both teams, and both phases. When Friedman tests were used to analyze differences among the four groups, significant differences were found for the data from the entire study and for team 2 but not for team 1 or for either phase. No results will be reported in the table below for those subsets of the data for which no significant differences were found.

Summary of Test Order Wilcoxon/Sign Results: Spanish Pilot Study						
	Fourth		Third		Second	
Overall Study						
First	Different		Mixed		Same	
Second	Mixed		Same			
Third	Same					
	Fourth		Third		Second	
	Team 1	Team 2	Team 1	Team 2	Team 1	Team 2
First	N/A-Same	Different	N/A-Same	Same	N/A-Same	Same
Second	N/A-Same	Mixed	N/A-Same	Same		
Third	N/A-Same	Same				
	Fourth		Third		Second	
	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2
First	N/A-Same	N/A-Same	N/A-Same	N/A-Same	N/A-Same	N/A-Same
Second	N/A-Same	N/A-Same	N/A-Same	N/A-Same		
Third	N/A-Same	N/A-Same				

This table reports a summary of the results of two non-parametric analyses of variance: the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In this case, the final negotiated ratings assigned to tests in order of administration as they were compared two at a time. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs,* ***different*** *indicates that both tests found a statistically significant difference between the pairs, and* ***mixed*** *indicates that the tests returned different results.* ***N/A-Same*** *indicates that results of a Friedman test showed that there were no statistically significant differences among the groups taken as a whole. The* ***overall*** *results take into account all of the tests administered during the English study; while the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively. In an ideal world, all of the pairs would have been found to be the same, with no statistically significant differences.*

The results of team 1, phase 1, and phase 2 are all marked as N/A-Same (Not Applicable) for the above table reporting results of Wilcoxon and Sign tests based on the Friedman results referred to above. These groups may be considered to have no statistical differences among them. For the data from the overall study and for team 2, the first and fourth tests seemed to be most different from one another. The second/fourth and first/third comparisons also indicated some differences

The correlations among the ratings for test order fell in the following pattern:

Summary of Test Order Correlation Results: Spanish Pilot Study			
Data Subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	.704	.751	.047
Team 1	.696	.774	.078
Team 2	.657	.810	.153
Phase 1	.745	.804	.059
Phase 2	.640	.739	.099

*This table reports the lowest and highest interagency Kendall's tau-b correlation coefficients for the final negotiated ratings assigned to tests in order of administration as they were compared two at a time. The column labeled **Δ** reports the difference between the two correlation columns. The **overall** results take into account all tests administered during the Spanish study; the **team 1** and **team 2** results take into account only those examinees tested by the set of testing pairs assigned to each team. **Phase 1** and **phase 2** results take into account those tests administered in the first and second halves of the study, respectively.*

These correlations indicate that the correlations for the ratings assigned during the second half of the study showed more differences than those assigned during the first half. They also seem to indicate that the correlation coefficients varied nearly twice as much for the team 2 pairs as they did for the team 1 pairs.

Summary of Test Order Median and Interquartile Range Results: Spanish Pilot Study				
Data Subset	**Low Median**	**High Median**	**Low Interquartile Range**	**High Interquartile Range**
Overall	2+	2+	10.0	10.0
Team 1	2+	2+	10.0	12.0
Team 2	2+	2+	10.0	10.0
Phase 1	2	2+	10.0	12.0
Phase 2	2+	2+	10.0	18.0

*This table reports the lowest and highest median and interquartile range calculated on the Spanish pilot study data on ratings assigned to tests in order of administration. The median is a measure of central tendency, and the interquartile range is a measure of the dispersion of the final ratings across the ILR scale. The **overall** results take into account all tests administered during the study; the **team 1** and **team 2** results take into account only those examinees tested by the set of testing pairs assigned to each team. **Phase 1** and **phase 2** results take into account those tests administered in the first and second halves of the study, respectively.*

In this study, no differences were shown between the teams except that the IQR is slightly higher for phase 2 than for the study overall or for phase 1. The results of these analyses show support for a trend identified previously—the variability of scores seemed to increase for phase 2 over phase 1.

BEST COPY AVAILABLE

Time of Administration. Timing effects were analyzed by grouping every examinee's 9:00 test in a single group, all 10:30 tests in a different group, and so on for their 1:00 and 2:30 tests. This section will report results on the level of agreement and differences between these groups for the study overall. Please see appendix E for additional details. Chi-squares tests found differences among all of the distributions for the individual time slots. A Friedman analysis of the four groups found no statistical differences among the different slot assignments nor when comparing all morning tests with all afternoon tests. Correlations ranged between .658 and .759, within a much narrower band than for the previously reported analyses.

English Pilot Study

The English study was set up to test the revised training materials to verify their usefulness in training novice testers and to determine what, if any, effect separate training workshops would have on the SPT results. The English study was smaller than the Spanish study, with half the number of testers conducting SPTs over six weeks. For this reason, no **team** analyses will be reported in the tables below; they will contain results for the study overall and for phase 1 and phase 2. The English testers were provided from USG organizations for training, and, unlike the Spanish testers who tested every other day, they administered four tests per day Monday through Thursday. Each test was videotaped and audiotaped, and the testing pair that conducted the test provided the live rating. A random selection of videotaped tests was re-rated by each of the other testing pairs in a specific pattern. The ratings of videotaped tests, referred to below as taped ratings, took place in circumstances as similar to the live ratings as possible, in that testers were asked to view each test in its entirety and provide a rating in one uninterrupted session, following the same rating procedures that they used to rate their own live tests. The taped ratings results were analyzed to provide additional information about test reliability among the four testing pairs.

The following questions will be addressed in the sections below:

- **Inter-pair reliability**:
  How well did the testing pairs agree on their final ratings for each examinee?
  How well did the live ratings from each testing pair agree with the taped ratings assigned by another pair?
- **Inter-rater reliability**:
  How well did the testers in each pair agree with one another on each test?
- **Effect on reliability caused by test order and time of administration**:
  Was there an effect on ratings caused by test order?
  Was there an effect on ratings caused by the time of day when the test was administered?
- **Effect on reliability caused by location of training**:
  Were there differences in ratings between the pairs trained on the west coast and those trained on the east coast?

Inter-Pair Reliability

This section will report on the results of analyses conducted to assess the amount of and patterns of inter-pair agreement and disagreement found among the final negotiated ratings for the English tests. One of the most important benefits and perhaps the main goal of this effort of creating and implementing a common speaking proficiency test is to ensure that a single examinee taking the new test will receive the same rating—no matter which testing pair administered the test. For this reason, it is expected that when the SPT is fully implemented, with joint training on a single set of test procedures, no significant differences will be found among the ratings by the different testing pairs. Cross-tabulation charts for the distribution of final ratings are included in this report as appendix H. The following tables provide data on how closely the English pilot data comes to this ideal.

Agency Rating Analyses. This section will report on the percent level of agreement among the testing pairs for each examinee for the overall study, for the first three weeks of data collection (phase 1), and for the second three weeks of data collection (phase 2). The results of the English pilot study from these rating analyses are included in the following tables.

Agency Rating Analyses—Exact Matches: English Pilot Study

	N	Exact Matches (4)	Exact Matches (3)	Exact Matches (none)
Overall	75	17 %	29 %	1 %
Phase 1[4]	41	22 %	37 %	0 %
Phase 2	34	12 %	21 %	3 %[5]

Exact Matches (4) *indicates the percentage of examinees for whom all testing pairs assigned exactly the same score.* ***Exact Matches (3)*** *indicates the percentage of examinees for whom at least three pairs assigned exactly the same score (including the percentage for whom all four agreed exactly).* ***Exact Matches (none)*** *indicates the percentage of examinees for whom all pairs assigned a different final score. The* ***overall*** *results take into account all tests administered during the English study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The percent level of exact agreement among four testing pairs varied slightly throughout the study. It also seems that there were more exact matches during the first half of the study than during the first half. These differences are not great, so they may reflect the nature of the examinees who participated in the two halves of the study rather than being due only to tester behavior.

---

[4] Two examinees (#142 and #189) in the English study performed their testing in different weeks so that their results split across the phases. For the purposes of these analyses, these examinees were counted within phase 1.

[5] Note that the percentages for overall and phase 2 represent a single examinee for whom none of the pairs agreed. The differences in the percentages are a function of the total number of examinees (N) included in the analyses.

The following table reports the percentage of examinees for which either four or three of the English testing pairs assigned ratings within the same base level.

Agency Rating Analyses—Within-Level Matches: English Pilot Study

	N	Within-Level Matches (4)	Within-Level Matches (3)
Overall	75	35%	64%
Phase 1	41	42%	76%
Phase 2	34	27%	50%

Within-Level Matches (4) *indicates the percentage of examinees for whom all testing pairs assigned exactly the same score plus those where all of the ratings fell within a given level, that is, where all four pairs assigned either a given ILR base level or its respective plus level (e.g., all 2 or 2+ ratings).* ***Within-Level Matches (3)*** *indicates the percentage of examinees for whom at least three pairs assigned scores within the same level (plus the percentage for whom all four agreed exactly, within-level, and when three agreed exactly). The* ***overall*** *results take into account all tests administered during the English study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

Within-level matches also seemed to be higher during the first phase of the study.

For the following tables, average percent level of exact agreement was calculated for each testing pair by comparing each pair's rating for each examinee with those assigned by each of the other participating pairs.

Agency Rating Analyses—
Percent Level of Agreement by Agency (Exact Matches):
English Pilot Study

	Pair 1	Pair 2	Pair 3	Pair 4	Average
Overall	41%	42%	42%	42%	42%
Phase 1	46%	50%	46%	51%	48%
Phase 2	35%	33%	39%	31%	35%

Ratings assigned to a given examinee by each pair were compared to those assigned by each of the other pairs individually, e.g., pair 1's percent level of agreement was calculated by averaging pair 1's percentage of agreement with pair 2, with pair 3, and with pair 4. The ***average*** *column reports the average for the overall study.* ***Exact matches*** *includes the percentage of examinees for whom the two agencies assigned exactly the same score. The* ***overall*** *results take into account all tests administered during the Spanish study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The level of agreement did not vary appreciably among the testing pairs, but the percentages seem lower for phase 2.

Similar analyses were conducted considering within-level agreement of one testing pair with every other participating pair.

Agency Rating Analyses—
Percent Level of Agreement by Agency (Within-Level):
English Pilot Study

	Pair 1	Pair 2	Pair 3	Pair 4	Average
Overall	55%	57%	57%	59%	57%
Phase 1	67%	68%	62%	69%	67%
Phase 2	52%	52%	54%	46%	51%

Ratings assigned to a given examinee by each pair were compared to those assigned by each of the other pairs individually, e.g., pair 1's percent level of agreement was calculated by averaging pair 1's percentage of agreement with pair 2, with pair 3, and with pair 4. The ***average*** *column reports the average for the overall study.* ***Within-level matches*** *includes the percentage of examinees for whom the two agencies assigned scores within the same level (plus the percentage for whom the pairs agreed exactly). The* ***overall*** *results take into account all tests administered during the Spanish study; the* ***team 1*** *and* ***team 2*** *results take into account only those examinees tested by the set of testing pairs assigned to each team.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

In a pattern similar to that observed for exact matches, the within-level results for phase 2 were lower than those for phase 1.

Statistical Analysis of Live Ratings. This section will examine the level of inter-pair agreement and differences among live final ratings. It will also provide results on the statistical similarities and differences and reliability coefficients found by various non-parametric analyses. Please see the English pilot data summary tables in appendix F for additional details.

The following table reports the inter-pair percent level of agreement for the English study.

Summary of Inter-Pair Percent Level of Agreement: English Pilot Study

	Low Percent-Agreement	High Percent-Agreement	Δ
Overall	41%	49%	8%
Phase 1	46%	58%	12%
Phase 2	30%	45%	15%

This table reports the lowest and highest percent level of agreement found among the comparisons made of the testing pairs in the English pilot study. The ***overall*** *results take into account all tests administered during the English study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively. In an ideal world, all of the pairs would have been found to have 100% agreement.*

The trend in the above analyses seems to continue, with percent-agreement higher during phase 1 than during phase 2.

The following tables report on the results of a number of non-parametric analyses. The non-parametric Pearson chi-squares analyses run to detect differences in how the ratings were

BEST COPY AVAILABLE

distributed across the scale by the four testing pairs showed that there were statistically significant differences among the four groups for the study overall and for both phases of the study. When a Friedman analysis was run comparing the four teams to one another, it indicated statistically significant differences among the groups. A significant Friedman result indicates that there are differences among the groups, but it does not identify where the differences can be found. Two additional tests, Wilcoxon and Sign, were run on each set of two pairs, comparing each pair to every other pair to determine where the differences were. Results from these two tests are reported in the table below.

Summary of Inter-Pair Wilcoxon/Sign Results: English Pilot Study						
	Pair 4		**Pair 3**		**Pair 2**	
Overall Study						
Pair 1	Different		Same		Same	
Pair 2	Different		Same			
Pair 3	Different					
	Pair 4		**Pair 3**		**Pair 2**	
	Phase 1	**Phase 2**	**Phase 1**	**Phase 2**	**Phase 1**	**Phase 2**
Pair 1	Mixed	Different	Same	Same	Same	Same
Pair 2	Different	Same	Same	Same		
Pair 3	Same	Different				

This table reports a summary of the results of two non-parametric analyses of variance: the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In this case, the final negotiated ratings assigned by the four testing pairs were compared two at a time. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs,* ***different*** *indicates that both tests found a statistically significant difference between the pairs, and* ***mixed*** *indicates that the tests returned different results. The* ***overall*** *results take into account all of the tests administered during the English study; the* ***phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively. In an ideal world, all of the pairs would have been found to be the same, with no statistically significant differences.*

As can be seen from the table above, the pattern of differences changed slightly depending upon the subset of the data being analyzed. These results indicate that the pairs 1, 2, and 3 did not differ from one another taking into account the data for the overall study and for both phases of the study. Pair 4 did show statistically significant differences from all of the other pairs, although the pattern of differences changed slightly from phase 1 to phase 2. This may reflect certain idiosyncrasies in pair 4. These results indicate that the pairs trained on the west coast rated similarly during the first half of the study (phase 1) but the pairs were statistically different during the second half of the study (phase 2) and for the study overall. The pairs trained on the east coast rated similarly throughout the entire data collection phase.

The following table reports on the correlations across testing pairs.

Summary of Inter-Pair Correlation Results: English Pilot Study			
Data Subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	.798	.853	.0550
Phase 1	.852	.910	.0580
Phase 2	.631	.866	.2350

This table reports the lowest and highest inter-pair Kendall's tau-b correlation coefficients for the final negotiated ratings assigned by the four testing pairs in the English pilot study as they were compared two at a time. The column labeled ***Δ*** *reports the difference between the two correlation columns. The* ***overall*** *results take into account all tests administered during the Spanish study;* ***phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The results of the Kendall's tau-b correlations indicate that the reliability among the groups dropped during the second half of the study, both in terms of raw correlations and in the range across which the correlations were spread.

The following data are related to the interquartile ranges around the various pair medians.

Summary of Inter-Pair Median and Interquartile Range Results: English Pilot Study				
Data Subset	**Low Median**	**High Median**	**Low Interquartile Range**	**High Interquartile Range**
Overall	2+	3	18.0	20.0
Phase 1	2+	2+	10.0	20.0
Phase 2	2+	3	10.0	12.0

This table reports the lowest and highest median and interquartile range calculated on the final ratings assigned by the four testing pairs during the English pilot study. The median is a measure of central tendency, and the interquartile range is a measure of the dispersion of the final ratings across the ILR scale. The ***overall*** *results take into account all tests administered during the study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The median scores and interquartile ranges did not vary much, perhaps due to the large number of examinees rated at levels 2+ and 3. The results indicate that the English ratings tended to vary from one plus level to a full base level up or down.

Statistical Analysis of Taped Ratings. For the English pilot study, a random sample of tests administered by each testing pair was re-rated by each of the other pairs, that is, pair 2 re-rated about eight videotaped tests administered by each of the other pairs (3, 4, and 5). Because the pattern of taped ratings for the English study was set up so that each testing pair rated a subset of tests administered by every other pair, no individual pair results could be reported, since each pair performed taped ratings on a different subset of examinees. However, results from all live ratings for the overall study will be compared to all taped ratings for the overall study. Please see appendix F for the details of the results. Statistical analyses were run comparing the live ratings to the taped ratings assigned to each test to determine the inter-pair reliability.

Percent Level of Agreement Comparing All Live Ratings to All Taped Ratings: English Pilot Study		
Percent Agreement	n=84	.4524

This table reports the percent level of agreement found by comparing the taped ratings for a given examinee to their respective live final ratings in the English pilot study. Each testing pair re-rated a different subset of examinees (made up of 6 to 9 examinees) as a part of the taped ratings, so individual agency percentages will not be reported due to the small sample sizes. In an ideal world, all of the pairs would have been found to have 100% agreement

The percent-agreement results for the comparison of live ratings with taped ratings fell within the same range as that for live ratings

The following paragraph summarizes the results of a number of non-parametric analyses of variance A Pearson chi-squares test comparing all taped ratings found a significant difference in the distribution of the scores on the scale However, Pearson chi-squares analyses run on the four English pairs' taped ratings yielded no significant differences The Friedman test indicated differences among the groups The following table reports the results of the Wilcoxon and Sign tests, showing where these differences were

Summary of Inter-Pair Wilcoxon/Sign Results—Taped Ratings Only: English Pilot Study			
	Pair 4	**Pair 3**	**Pair 2**
Pair 1	Different	Same	Mixed
Pair 2	Same	Same	
Pair 3	Different		

This table reports a summary of the results of two non-parametric analyses of variance the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In this case, all the taped final ratings assigned by the four testing pairs were compared to one another. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs,* ***different*** *indicates that both tests found a statistically significant difference between the pairs, and* ***mixed*** *indicates that the tests returned different results In an ideal world, all of the pairs would have been found to be the same, with no statistically significant differences.*

Pair 2 seemed to behave differently in the taped ratings than it did in live ratings, showing similarities with pair 4 and differences with pair 1

Summary of Inter-Pair Correlation Results—Taped Ratings Only: English Pilot Study			
Data subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	623	1.000	.3770

This table reports the highest and lowest inter-pair Kendall's tau-b correlations for the taped ratings from the English study. The column labeled Δ reports the difference between the two correlation columns. The ***overall*** *results take into account all tests administered during the study*

The correlations comparing the agencies taped ratings to one another were lower than the reliability levels for live ratings, ranging from .623 to 1.000. These differences may be tied to the lower number of examinees taken into account in these results (about 24) than for the live ratings (about 125). As was mentioned above, tests were run only for the study overall rather than for the subsets of the data reported for live ratings because there were not enough taped ratings performed to divide the data set further.

Inter-Rater Reliability

This section will examine the level of within-pair agreement in individual final ratings for live ratings within the English study.

Statistical Analysis of Live Ratings. Reliability results will be reported in terms of percent level of agreement as well as correlations for each tester's individual rating with his or her testing partner for the study overall and for phase 1 and phase 2. Please see appendix F for the summary results tables on the English pilot study.

Summary of Inter-Rater Percent Level of Agreement: English Pilot Study					
	Pair 1	**Pair 2**	**Pair 3**	**Pair 4**	**Average**
Overall Study	68%	51%	61%	91%	68%
Phase 1	71%	56%	58%	93%	70%
Phase 2	65%	44%	65%	88%	66%

*This table reports the percent level of agreement between live individual tester ratings within each testing pair in the English pilot study. The **overall** results take into account all tests administered during the English study. **Phase 1** and **phase 2** results take into account those tests administered in the first and second halves of the study, respectively. In an ideal world, all of the pairs would have been found to have 100% agreement.*

In general, phase 1 percent-agreement was higher than that of phase 2, although the behavior of the pairs varied individually between phases.

The following table reports the correlation coefficients for inter-rater reliability.

Summary of Inter-Rater Correlation Results: English Pilot Study			
Data subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	.845	.966	.1210
Phase 1	.864	.984	.1200
Phase 2	.738	.936	.1980

*This table reports the lowest and highest interagency Kendall's tau-b correlation coefficients for the live individual tester ratings within testing pairs in the English pilot study. The column labeled **Δ** reports the difference between the two correlation columns. The **overall** results take into account all tests administered during the Spanish study; **phase 1** and **phase 2** results take into account those tests administered in the first and second halves of the study, respectively.*

Correlations varied within about the same range for phase 1 as for the overall study. However, phase 2 results show greater variance as well as lower agreement. These results indicate that raters tended to disagree more often during the second half of the study than during the first half. This trend is opposite that observed in the Spanish results in that the Spanish testers' inter-rater agreement increased in phase 2.

Statistical Analysis of Taped Ratings. Inter-rater statistical analyses were also run on the ratings assigned by testers during the taped ratings. The following table contains data on the correlations for individual tester ratings on taped ratings only.

Summary of Inter-Rater Correlation Results—Taped Ratings Only: English Pilot Study		
Lowest Correlation	**Highest Correlation**	**Δ**
.829	1.000	.1710

This table reports the percent level of agreement and lowest and highest Kendall tau-b correlation coefficients between taped individual tester ratings in the English pilot study. The column labeled Δ reports the difference between the two correlation columns.

Correlations were slightly lower for taped ratings than for the live ratings. This may be due to the reduced number of examinees (24) included in these analyses compared to that of the live ratings (125).

Effects on Reliability by Test Order/Time of Administration

This section will report on the results of analyses conducted to assess the amount of agreement and disagreement found among the final negotiated ratings for the English tests examining the tests by test order and time of administration. The data collection schedule was designed to counterbalance for variance due to test order effect and timing effects, including examinee practice effect, examinee fatigue, and tester fatigue. It was expected that examinees would become better at performing the different sections of the SPT with multiple administrations and that perhaps their scores would improve slightly, but it was also believed that the act of taking two tests one after another would tire examinees and reduce their scores slightly. Members of the FLTB were concerned about potential effects from tester fatigue as well, in that the English study's schedule called for each individual tester to administer many more tests per week than was called for in the Spanish study schedule. The research design of the study was counterbalanced to control for these effects by spreading it across all examinees; however, the results from this section may be of interest to program managers as they arrange testing schedules.

Test Order Effects. Test order effects were analyzed by grouping the final negotiated rating for every examinee's first test in a single group, all second tests in a different group, and so on for their third and fourth tests. This section will examine the level of agreement and differences between testing pairs for the study overall and for both phases. It will also provide reliability coefficients for the same groups. This section will take into account live ratings only. Please see the English pilot data summary tables in appendix F for results of non-parametric chi-squares tests, non-parametric analyses of variance for the patterns of agreement between test order groups, and correlations on the four final ratings for the reliability coefficients.

Pearson chi-squares tests run on the distribution of rating for first, second, third, and fourth tests showed that there were statistically significant differences among all of the groups for the study overall as well as for both phases. When Friedman tests were used to analyze differences among the four groups, no significant differences were found for the overall study or for either phase. The group made up of every examinee's fourth test showed some differences from the other tests for the overall study, but these differences were not statistically significant. For phase one, the fourth test results only differed from those of the first and second tests; for phase 2, the fourth test results differed from the first and third test results.

Summary of Test Order Correlation Results: English Pilot Study			
Data Subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	.756	.821	.0650
Phase 1	.813	.874	.0610
Phase 2	.587	.727	.1400

This table reports the lowest and highest interagency Kendall's tau-b correlation coefficients for the final negotiated ratings assigned to tests in order of administration as they were compared two at a time. The column labeled ***Δ*** *reports the difference between the two correlation columns. The* ***overall*** *results take into account all tests administered during the Spanish study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

These correlations show much higher agreement among the ratings during the first half of the study than for the second half.

Summary of Test Order Median and Interquartile Range Results: English Pilot Study		
Test Order	**Median**	**Interquartile Range**
First	3	18.0
Second	2+	18.0
Third	2+	18.0
Fourth	3	18.0

This table reports the medians and interquartile ranges calculated on the final ratings assigned to the tests in test order by testing pairs in the English pilot study. The median is a measure of central tendency, and the interquartile range is a measure of the dispersion of the final ratings across the ILR scale. Results are presented for the ***overall*** *study only.*

The medians and interquartile ranges for the test order groups support the Friedman results; there are no significant differences.

Time of Administration. Timing effects were analyzed by grouping every examinee's 9:00 test in a single group, all 10:30 tests in a different group, and so on for their 1:00 and 2:30 tests. Please see the English pilot data summary tables in appendix F for additional details. The non-parametric chi-squares test found differences in the distributions of ratings for each of the groups. Friedman results that compared the test slot groups to one another approached significance ($p = .0519$). In the Wilcoxon and Sign tests, the 9:00 a.m. tests tended to differ from the other tests. A Friedman test run comparing all of the morning tests (9:00 a.m. and 10:30 a.m.) with all of the afternoon tests (1:00 p.m. and 2:30 p.m.) showed no significant differences among the groups. Correlations ranged from .794 to .863, with the lowest correlation between the 1:00 and 2:30 tests.

Effect on Reliability by Location of Training

As was mentioned previously, the English pilot study was designed to answer slightly different research questions than the Spanish pilot study. One of these differences relates to evidence of rating variance caused by separate tester training workshops. The following tables will report differences in ratings between the pairs trained on the west coast and those trained on the east coast.

The following table contains the inter-group percent level of agreement between the two training groups in the English study.

Summary of Training Group Inter-Pair Percent Level of Agreement: English Pilot Study	
Overall Study	42%
Phase 1	48%
Phase 2	35%

*This table reports the percent level of agreement between final ratings assigned by testing pairs trained on the east coast with those assigned by the pairs trained on the west coast. The **overall** results take into account all tests administered during the English study. **Phase 1** and **phase 2** results take into account those tests administered in the first and second halves of the study, respectively.*

The percent-agreement seems to be slightly higher in phase 1 than in phase 2.

Summary of Training Group Inter-Pair Correlation Results: English Pilot Study	
Data Subset	**Correlation**
Overall	.724
Phase 1	.758
Phase 2	.652

*This table reports the Kendall's tau-b correlation coefficients comparing the final ratings assigned by testing pairs trained on the east coast with those assigned by the pairs trained on the west coast. The **overall** results take into account all tests administered during the Spanish study. **Phase 1** and **phase 2** results take into account those tests administered in the first and second halves of the study, respectively.*

As was shown previously, phase 1 showed higher correlation coefficients than phase 2 as well.

Training Group Inter-Pair Median and Interquartile Range Results: English Pilot Study		
Data Subset	**Median**	**Interquartile Range**
East Coast Training Workshop		
Overall	3	18.0
Phase 1	2+	18.0
Phase 2	3	11.0
West Coast Training Workshop		
Overall	2+	18.0
Phase 1	2+	20.0
Phase 2	3	12.0

This table reports the medians and interquartile ranges calculated on the final ratings for the English pilot study data for the two training groups. The median is a measure of central tendency, and the interquartile range is a measure of the dispersion of the final ratings across the ILR scale. The ***overall*** *results take into account all tests administered during the study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

The interquartile range is greater for phase 1 than for phase 2 for both groups indicating that the ratings were spread more widely. This result could be due to characteristics of the examinees tested in the different phases rather than due strictly to tester behavior. Overall, the east coast ratings seemed to be slightly higher, and the west coast ratings were dispersed across a wider range.

The following table refers to results of a number of non-parametric analyses of variance. No chi-squares or Friedman analyses were run on the training group data. Only Wilcoxon and Sign results will be reported in the following table.

Summary of Training Group Inter-Pair Wilcoxon/Sign Results: English Pilot Study	
Overall	Different
Phase 1	Different
Phase 2	Same

This table reports a summary of the results of two non-parametric analyses of variance: the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In this case, the final ratings assigned by the two training groups were compared. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs;* ***different*** *indicates that both tests found a statistically significant difference between the pairs. The* ***overall*** *results take into account all tests administered during the study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively. In an ideal world, all of the pairs would have been found to be the same, with no statistically significant differences.*

These results indicate that over time, the testers in the training groups became more similar, perhaps due to the increased variability of the ratings. The groups were statistically significantly different for the study overall and for phase 1.

The following table contains the inter-rater percent level of agreement for the English study training groups.

Summary of Training Group Inter-Rater Percent Level of Agreement: English Pilot Study		
	East Coast	**West Coast**
Overall Study	59%	76%
Phase 1	63%	75%
Phase 2	58%	77%

This table reports the percent level of agreement between live individual tester ratings in the English pilot study. The ***overall*** *results take into account all tests administered during the English study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively. In an ideal world, all of the pairs would have been found to have 100% agreement.*

The inter-rater percent level of agreement for the east coast testers seemed slightly lower than for those of the west coast. The inter-pair percent-agreement levels were about 20% lower than these inter-rater results.

The following table contains the correlations coefficients for inter-rater reliability.

Summary of Training Group Inter-Rater Correlation Results: English Pilot Study			
Data Subset	**East Coast Group**	**West Coast Group**	**Δ**
Overall	.860	.917	.0570
Phase 1	.900	.917	.0170
Phase 2	.761	.905	.1440

This table reports the Kendall's tau-b correlation coefficients comparing the final ratings assigned by the two training groups. The column labeled ***Δ*** *reports the difference between the two correlation columns. The* ***overall*** *results take into account all tests administered during the English study.* ***Phase 1*** *and* ***phase 2*** *results take into account those tests administered in the first and second halves of the study, respectively.*

Inter-rater reliability is well above the 70 % cut-off level selected for the study for both groups, although the results are slightly lower for the east coast group.

Section 7. SPT Validity

The pilot validation studies were designed to address a number of questions related to the validity of the new SPT testing procedures.

Current Thought on the Notion of Validity

In recent years, there has been an explosion of scholarly discussion about test validity, and a number of assumptions about validity have been debated. Scholars still tend to agree on a basic definition of the term: validity measures truth in testing; i.e., whether and to what extent a test does truly measure what it purports to measure. However, there has been active revision of the conceptual notion. Validity is no longer thought to be comprised of separable types of validity, e.g., "construct validity," "criterion-related validity," "face validity"—each with their separate measures. Rather, validity is seen as a unitary concept, and what were formerly known as 'types of validity' are seen more as 'sources of evidence for validity,' each of which has the potential to contribute to test validation. At present, the international language testing community is still engaged in ongoing debate on this issue, but it is clear that the unitary definition of validity is now pretty well entrenched. The discussion of SPT validity in this section is presented in the framework of unitary validity.

Literature Review: A Unified Concept of Validity

A excellent summary of recent changes in test validity can be found in the following reference:

Gronlund, N. E. (1993). "Validity and Reliability." *How to Make Achievement Tests and Assessments*. Boston: Allyn and Bacon.

The seminal reformulation of validity in modern scholarship appears in a long paper by Samuel Messick. This paper does much more than argue for a unified definition of validity; for example, it reasons that modern validity must attend to the consequences of test usage—the impact a result has on the examinee's career and life. The citation for this article is the following:

Messick, S. (1989). "Validity." In R.L. Linn (Ed.), *Educational Measurement*, 3rd. Edition. New York: ACE/MacMillan.

Arguably, consequential evidence of validity is synonymous with the notion of washback. For an important re-evaluation of washback (the effect of testing on learning and teaching), see the following article:

Alderson, J.C. and D. Wall. (1993). "Does Washback Exist?" *Applied Linguistics* 14.2, 115-129.

There are unresolved questions about the new, unified concept of validity and its practical implications on test validation procedures. For a provocative discussion in that regard, see the following article:

Shepard, L. A. (1993). "Evaluating Test Validity." In L. Darling-Hammond (Ed.) *Review of Research in Education 19*. Washington, DC: AERA, pp. 405-450.

Finally, the revision of the definition of validity has implications for the establishment of test reliability, often thought to be a necessary (but not sufficient) precondition to test validation. A particularly stimulating modern treatment of this relationship can be found in the following article:

Moss, P. A. (1994). "Can There Be Validity Without Reliability?" *Educational Researcher* 23(2), pp. 4-12.

Although a clear process for validating a language test has yet to be developed and agreed upon by the international language testing community, one clear chord resonates through the scholarly discussions. Test validity no longer can be based upon only certain statistical measures of a few types of validity. A valid test must now demonstrate multiple sources of evidence of validity, where each source is of equal prestige in crafting an argument that the test is valid. With that in mind, the following is a discussion of the validity of the SPT using such multiple sources of evidence.

Convergent Evidence of Validity

Convergent evidence of validity includes evidence that shows improvement in results over the course of multiple pilot studies. The data thus far from the Spanish and English studies indicate that there has been an increase, in both studies, over the level of agreement among the various testing pairs participating in the 1986 study. This pattern of increased agreement for various statistical analyses indicates that the new common procedures introduced with the SPT appear to provide a more consistent rating across testing pairs than when agencies use their own testing procedures. These increases have been particularly noticeable in the data related to inter-pair agreement, where the percentages of exact and within-level agreement among testing pairs have increased, and where the percentage of cases where all four teams assigned different scores has decreased.

Concurrent Evidence of Validity

Concurrent evidence of validity for a new test is based on evidence of assignment of equivalent scores to a given examinee on other tests for which reliability and validity have been established and which arguably measure the same underlying trait. Recognizing that each examinee's previous OPI score as certified by the various agencies would provide proper concurrent evidence, CALL requested this information from the FLTB agencies. During the test development process, FLTB agencies were surveyed to determine whether it would be possible to release previous OPI scores for those government employees who volunteered as examinees.

Concerns related to privacy, freedom of information, and security were raised. In response to these concerns, the various FLTB agencies who sent their employees to participate in the pilot studies did not provide this data to CALL. Instead, examinees were asked to provide their most recent OPI Speaking scores in a pre-test questionnaire. Because providing this information was voluntary, a number of examinees chose not to provide this information.

In the Spanish study, a little over half of the examinees (53.6%) reported previous OPI Spanish results. The correlations of these previous OPI scores with the four final ratings assigned during the Spanish study ranged from .77 to .83, a relatively strong relationship. However, the following elements of the data collection process should be taken into consideration when evaluating this correlation. The first concern is that some examinees voluntarily provided the information, while others did not, and this group of examinees who chose to report may not be characteristic of the larger population. A second concern is related to differences among current OPI procedures. The examinees reported scores from tests administered at a number of agencies, including CIA, DLI, FBI, FSI, as well as the Peace Corps and from universities. Because these tests vary slightly in their test format, it is possible that these differences could have an effect on the scores. Another concern is related to the age of the scores. The examinees reported scores that ranged from 4 months to 16 years old. Research in the area of language attrition shows that proficiency often changes over time. The pattern of change depends on a large number of individual and environmental factors that could introduce differences in these scores, either for better or worse.

In the English study, examinees were also asked to provide their most recent OPI Speaking score in a pre-test questionnaire. The total number of English pilot study participants who reported prior oral proficiency ratings was too low to calculate any correlation between past scores and SPT scores.

Face Evidence of Validity

In the new, unified conception of validity, all sources of evidence are important, where previously, certain validity types might have been given more weight than others. Face evidence of validity includes the perceptions of the test by examinees and test administrators. Feedback from examinees has been collected during both pilot studies, and a summary of some of their responses related most closely to perceptions of the SPT's validity is included in the section below. Testers were also asked to provide extensive feedback in regular tester meetings with CALL staff and FLTB members and tester trainers as well as on a written survey. A summary of positive support for the new SPT procedures provided by the testers is included in this section as well.

Examinee Feedback

In the post-test questionnaires filled out by each examinee during the Spanish and English pilots, the examinees were asked the following two questions: (a) do you feel that the testers heard a good sample of the Spanish/English you know?, and (b) do you feel the testers found the limits of your language ability?

BEST COPY AVAILABLE

The results from examinee questionnaires are summarized in the table below.

Examinee Feedback Results

	Spanish				English			
	Good Sample		**Found Limits**		**Good Sample**		**Found Limits**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
YES	229	93%	206	87%	284	92%	227	77%
NO / other	16	7%	30	13%	26	8%	68	23%
Total	245		236		310		295	

Note: In the Spanish study, each examinee answered these questions twice, once after the second and once after the fourth test. In the Spanish study responses, there were cases in which examinees reported that they were challenged in one test, but not in another. This kind of response was coded in the data above as a "no." In the English study, the examinees responded after each test for a possible total of four responses per examinee.

In both studies, over 90% of the responses indicate that, in the examinee's opinion, the SPT elicited a good sample of their actual language ability. As it is difficult to measure real-life use of the language, the examinees' responses to this question provide useful subjective evidence for the validity of the SPT.

The percentages of responses that indicate examinees felt challenged to the limits of their ability are high (77 to 87%). It appeared that examinees had different interpretations of what was meant by "being challenged to the limits of their ability." Some answered "no" to the question of whether or not they had been challenged, then expanded on their answer by saying, "no, they were not challenged the whole time." It should be noted that the SPT procedures are such that SPT testers should not conduct the *entire* test at a level highly challenging to the examinee. To find the examinee's limits, several instances of breakdown must be shown. Also, breakdown may not be recognizable to the examinee, particularly to examinees at higher levels. Examinees often indicated that they were able to "talk around" their weak areas, and, thus, because they could avoid being driven to silence, did not feel their limits were reached. The testers, on the other hand, as native speakers of the language and through their training in the SPT procedures, might have become aware of subtle forms of breakdown that the examinee would be unable to notice, such as structural mistakes or non-native-like speech produced by the examinee.

Tester Feedback

At the end of both pilot studies, each of the participating testers was asked to fill out tester questionnaires to provide feedback on their experiences. The section below contains a summary of their perceptions on the validity of the different test sections.

Conversation. Testers often commented that the three-way conversation was very natural and that the conversation portion of the test, by helping to put examinees at ease, enabled examinees to converse more naturally.

Situation. Testers reported that the situation provided useful information on what examinees could do in practical, real-world situations. Through situations, different kinds of vocabulary could be more easily tested, speech contexts other than polite, informal conversation could be explored, and an array of tasks could be accomplished that were difficult to accomplish during conversation.

Information Gathering Task. Tester felt that the IGT was an effective way to assess interactive comprehension and that it added useful information about the examinee's ability to ask questions and about his or her communication strategies.

Content Evidence of Validity

Content evidence of validity includes evidence that the elements of the test are representative of the content area or context in which the examinee will function. The decision as to how representative these elements are is derived from the process of consensus-building undertaken by the test developers, in this case, by the FLTB. The FLTB dealt with a number of issues related to content validity during the test development phase, and, as a result of these discussions, the FLTB expanded and refined the definition of a ratable sample to increase the content validity of the SPT. As these discussions evolved, so too did the the FLTB's understanding of the traits measured in the SPT. This evolution of a common understanding is strong content evidence of validity as captured in records such as meeting minutes and the revised *Test Specifications* document (see Lynch and Davidson, 1994).

For example, the FLTB created a new set of rating factors for use by SPT raters. The process by which these factors were defined was built upon FLTB discussions of the contexts in which the SPT would be used and analysis of the ILR Speaking Skill Level Descriptions. The consensus of the FLTB members was that the ILR Skill Level Descriptions provide holistic descriptions of examinee proficiency and that, as such, holistic rating should be emphasized over separate ratings for the given factors. It was recognized, however, that testers seemed to benefit from use of such factors, which break down the examinee's language performance into various linguistic categories, such as grammar or vocabulary, both during tester training and during actual test administration. This set of factors now differs slightly from those in use at any of the FLTB agencies, particularly the Interactive Comprehension and Communication Strategies factors. These two factors in particular were included by the FLTB for the backwash effect it was hoped that they would have on teaching at the various agencies.

Another example of consensus building that can be considered as content evidence of validity took place when the FLTB was developing the procedures for the IGT as an approved SPT elicitation technique. The decision to include Interactive Comprehension as a factor created the need for discussions about how testers could verify an examinee's Interactive Comprehension during the SPT. The Board attempted to balance concerns about introducing English into the SPT due to stresses such codeswitching might place on examinees and testers with other concerns about failing to detect whether examinees truly comprehended the information they had collected.

The IGT was piloted with different variations, and the FLTB came to the agreement that the examinee would generally report back what he or she had learned in English to maximize the testers' ability to verify each examinee's Interactive Comprehension.

During the Spanish study, both testers remained in the room during the IGT; however, during the English study, since the entire process, including the report, was conducted in English, one tester consistently left the room. This decision provoked further discussion. The FLTB searched as a group to find a balance between concerns that rating reliability could be affected since one tester missed part of the examinee's sample and concerns about authenticity of the task since examinees and testers alike found the situation difficult to believe when one tester had to pretend not to have heard what happened.

Another key source of content evidence of validity is the relationship of the SPT to the ILR Skill Level Descriptions for Speaking. These criteria, developed in their earliest forms at the Foreign Service Institute just after WWII and refined by work at FSI and other language testing organizations since that time, have been adopted for over 15 years by all USG organizations. The FLTB's earliest discussions centered on whether to base the new speaking test on the ILR Skill Level Descriptions. The Board debated alternatives, and consensus was reached that the ILR Descriptions should be retained as the criteria for rating the new test because they been accepted as proficiency test criteria across the USG for 15 or more years. This process reaffirmed the ILR Descriptions as the basis for proficiency testing under the ULTP.

After consensus was reached that the Descriptions were indeed to be used, the FLTB, during its discussions and work in the test development phase, specifically built in processes by which testers are required to return to the actual wording of the ILR Descriptions during elicitation and rating. Testers currently administering speaking tests in the various FLTB agencies refer to agency-specific testing and rating aids containing summaries of the ILR descriptions while administering tests. In many cases, the wording of these aids is different from the wording in the ILR descriptions. During the design phase for the SPT tester rating and elicitation aids, the FLTB deliberately included the exact wording of the ILR descriptions and carefully marked any additional wording that did not come from the descriptions to distinguish it from the original. In particular, the wording on the *Elicitation Aid* extracts wording from the descriptions of elements significant at each level, and the *Rating Factor Grid* is a reorganization of phrases from the ILR Skill Level Descriptions that are associated with each of the six rating factors.

Another refinement (which stands as content evidence of validity) was the FLTB's conceptualization of plus levels and their treatment within the rating process. The ILR descriptions indicate that each of the six base levels is a threshold for a level. Plus levels were added for levels 0 through 4, increasing the number of points on the scale to 11. The FLTB decided to treat plus levels as the uppermost area of base level ranges rather than as separate ranges in themselves. Because of this decision, SPT testers now identify the base level that best describes an examinee's performance first, and then, as a second step, testers compare the description of the original base level with that of its related plus level to determine which is most appropriate. The decision to treat plus levels in this way strengthens the scale by asking raters to discriminate first among six base levels rather than simultaneously among six base levels and five plus levels.

As the FLTB work progressed, it became clear that each participating agency had its own cultural assumptions about oral testing in the federal government. One key benefit of the SPT—a type of 'backwash' perhaps—has been to foster extremely focused discussions of the design and uses of oral testing procedures in very different government contexts. This increased communication and collaboration has been a rewarding byproduct of the FLTB's work on the ULTP, and it is hoped that such communication will continue and improve as the project continues through SPT implementation and in the next stages on Listening, Reading, and Writing.

Validity: Concluding Remarks

This section has reviewed aspects and results from the first two Speaking Proficiency Test validation studies related to validity, taking into consideration recent reformulations of validity in the scholarly literature. Evidence has been presented in the form of empirical data, reports of theoretical development and conceptualization of the new procedures, reports of tester and examinee reactions to the new test, and a discussion of the evolution of the SPT design. These conclusions provide strong support for the validity of SPT scores as measures of overall speaking proficiency, as required by USG personnel in their daily activities.

Section 8. Recommendations

Recommendation 1: Based on the positive results from the Spanish and English studies and preliminary results from the Russian study, begin moving toward pilot operational implementation at the various agencies, resources permitting.

Discussion: There has been an increase, maintained in both SPT pilot studies, in the level of agreement among the various testing teams over those reported in the 1986 study. This pattern of increased agreement over the last study of this type indicates that the new common testing and rating procedures introduced with the SPT appear to provide for more consistent rating across teams from different agencies than did the use of agency-specific procedures.

In addition, there is now evidence that the SPT elicitation and rating procedures are robust, sufficient, and appropriate, even under the following conditions:

- When testers are trained separately in two sites (as in the English study) by two interagency teams of trainers using the SPT training curriculum and methodology.
- When newly trained testers are used (as in the English study).
- When experienced testers are retrained (as in the Spanish study).

It is evident that administration of the SPT in the two pilot studies resulted in much greater agreement among teams than the levels found in the 1986 studies. The greater reliability attained with the Spanish pilot study over that level found in the English pilot study is also interesting. Even with novice testers, there was evidence of increased reliability in the English study over that of the 1986 study, although the results for the Spanish study that used experienced testers were slightly higher. These increases in reliability reflects the greater homogeneity achieved among the testing teams with the latest implementation of the SPT training methodology.

Some Board member agencies have decided to conduct pilot operational implementations in the field, where testers in specific languages will be retrained in the SPT procedures. In addition, some agencies are planning small-scale comparability studies of the results on the new SPT with those of their current test on the same examinees. The results from ongoing reliability studies carried out during these pilot implementations and comparability studies will provide further information on specific conditions at the various agencies and with various languages.

Recommendation 2: Contingent upon a positive outcome to the Russian pilot study and pilot operational implementations and upon individual agency approval, fully implement the SPT in all languages.

Discussion: Given the level of agreement among the agencies participating in the Spanish and English studies and the positive results from tests of various aspects of the SPT methodology, agency managers may realistically begin moving toward full implementation of the interagency SPT program for oral proficiency assessment. Assuming continued positive results from the Russian pilot validation study and pilot operational implementation projects, the newly developed SPT instrument procedures and training materials are ready for operational implementation. Once all data are collected and analyzed, each agency will receive a recommendation report from the FLTB about adopting the new SPT procedures.

Efforts and resources should now be devoted to identifying agency-specific requirements for implementation of the new test procedures, including resource requirements, and to strengthening other aspects of the SPT's operational use and implementation, rather than to the planning and implementation of further pilot studies. For example, the Chinese pilot study originally planned for next year is canceled.

Recommendation 3: Maintain ongoing interagency collaboration on language proficiency testing.

Discussion: It is of great importance at this time to continue firm interagency coordination and collaboration on testing procedures. The collaboration brought about through the FLTB's efforts should continue, not only in implementing the ULTP for other language skills, but in the development and maintenance of quality control procedures that allow for the monitoring of testing activities to ensure that common standards are maintained. In this process of interagency quality control and accountability, CALL could play a very useful role by providing a setting and the necessary technical, professional, and funding support for the interchange of data and the analysis and interpretation of results. This mechanism is of primary importance to maintain the level of interagency agreement and mutual collaboration that has been achieved under the ULTP.

Recommendation 4: Coordinate interagency work on pilot operational implementation projects and quality control procedures during pilot and full implementation.

Discussion: An issue that has strong implications for the ULTP and which impacts on all matters previously discussed is the continued coordination of testing under the ULTP to maintain and even improve on the progress achieved with the development of the SPT to date. Specifically, it is necessary to establish a process of periodic, systematic, and planned data analysis and interpretation. This process should involve some form of random sampling of the test data obtained at the various agencies participating in the SPT implementation and obtaining new ratings of the taped tests by the other agencies. The various statistical analyses that can be performed on the data will establish if the agencies remain within acceptable ranges of agreement in their rating outcomes. This process should include the analysis of data from studies designed to examine the effects of other modes of testing, such as telephone- and video-teletesting (VTT), when using the SPT methodology.

The reliability program should be centralized, and the FLTB should be the organization that designs and, through CALL, implements this plan and the specific form of data sampling, and which reports periodically to the agencies on the results of the analyses so that any necessary corrective action can be taken. Remedies for potential rating drift among agencies could include retraining seminars for testers or review of testing practices, among other activities.

Analyses should also be conducted on the testing behavior and rating results of individual testing pairs and testers within the agencies. As a result of these studies, corrective measures could be suggested for testers who do not reach the expected levels of reliability in their ratings. These measures could include retraining seminars and testing under supervision. It would be essential, though, that if, after all attempts at remediation have been taken, the quality of testing performance of certain testers does not reach the minimum acceptable standards of reliability, they should be reassigned to functions other than testing. This process will ensure the continued quality and comparability of interagency scores. In addition, it would give management within each FLTB agency the necessary reliable information and confidence in the language proficiency reports.

These studies and analyses should specifically observe the percent level of agreement between the ratings produced by different tester teams and/or agencies, particularly regarding exact matches and within-level matches in the ratings.

Similarly, the process of interagency collaboration that proved essential to the success of the ULTP projects undertaken thus far should continue with frequent exchange of opinions, workshops, training, and other activities. In this manner, the successful interagency collaboration will continue and will undoubtedly produce further progress.

Recommendation 5: Consider adjusting the format of SPT tester training workshops, based on results of these studies that indicate that retraining of experienced testers requires more time than originally expected. Additional time may be dedicated to formal classroom learning, practice testing, and apprenticeships in addition to possible self-study components.

Discussion: Improved results in terms of greater reliability have seemed to increase in relation to the amount and focus of training provided to the testers involved in the studies. The new increase in level of agreement between teams has been accompanied by a significant increase in training and practice time. The Spanish and English studies involved a great increase in training time over the average training currently in place in the language testing community. At the same time, they showed a remarkable improvement in interagency and inter-pair percent-agreement over the 1986 (and probably current) levels of agreement among testers using agency-specific tests. Tester training for the English pilot study benefited from more fully developed training materials and a more structured syllabus.

Although more aspects of this correlation between increased reliability of the scores and training time should be explored, it seems that they are highly correlated, and it is important for testing program managers to consider the feasibility of planning more extensive training (and retraining) sessions for language testers. The experience during the pilot studies suggests that a period of two weeks seems adequate for the more formal aspect of training. In addition, an extended period of some form of guided practice with feedback from experienced trainers is probably very important to achieve the level of training that would guarantee the quality of testing and rating practices needed in the USG.

Given the operational and budgetary constraints in the participating agencies, it may be appropriate that the two weeks of standard formal training under the SPT procedures for new testers be complemented by an extended period of "apprenticeship." If it is not possible to provide a period of intensive "practice training" immediately after the two weeks of classroom training, a formal system should be created in which new testers administer tests operationally under the extended supervision of experienced testers for a period of six months or more. Ratings by the apprentice testers should perhaps not carry the same weight as those of fully certified testers. A tester would not be considered fully certified until the person has achieved the levels of reliability desired according to interagency procedures. Intensive and extended training with objective "exit" assessment is necessary to achieve the quality program desired and to improve further the reliability of ratings currently achieved in the pilot studies.

Section 9. Bibliography

Alderson, J.C. and D. Wall. (1993). "Does Washback Exist?" *Applied Linguistics* 14:2, 115-129.

Armstrong, M., I. Cornwell, M. Fogel, K. Glasgow, M. Johnson, A. Kellogg, I. Knippler. (1992). *Proposal for the Language Proficiency Testing Board*. Unpublished manuscript. Arlington, VA: Center for the Advancement of Language Learning.

Clapham, C.M. (1994). *The Effect of Background Knowledge on EAP Reading Test Performance*. Unpublished dissertation. Lancaster University: Lancaster, UK.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement*. 20:1, 37-46.

Cole, D.J., P.E. Messner, H. Swonigan and B. Tillman. (1991). *Portfolio Structure and Student Profiles: An Analysis of Education Student Portfolio Reflectivity Scores*. Paper presented at the AERA national conference. ERIC Document Reproduction Service (microfiche) number ED 335 307.

Everitt, B.S. (1968). "Moments of the Statistics Kappa and Weighted Kappa." *The British Journal of Mathematical and Statistical Psychology*. 21: 97-103.

Fleiss, J.L. (1971). "Measuring Nominal Scale Agreement Among Many Raters." *Psychological Bulletin*. 75: 378-387.

Gronlund, N. E. (1993). "Validity and Reliability." *How to Make Achievement Tests and Assessments*. Boston: Allyn and Bacon.

Hart-Gonzalez, L. (1993). *The Role of Proficiency Testing in Federal Research*. Paper presented at the annual RP-ALLA Conference, Ohio State University.

Hatch, E. and A. Lazaraton. (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. New York: Newbury House Publishers.

Kaplan, B.A. and E.G. Johnson. (1992). *Reliability of Professionally Scored Data: NAEP-Related Issues*. Paper presented at the AERA Annual Meeting. ERIC Document Reproduction Service (microfiche) number ED 347 186.

Lynch, B. and F. Davidson. (1994). "Criterion-Referenced Language Test Development: Linking Curricula, Teachers, and Tests." *TESOL Quarterly*. 28:4, pp. 727-743.

Mason, E.M. (1992). "Percent of Agreement Among Raters and Rater Reliability of the Copying Subtest of the Stanford-Binet Intelligence Scale." Fourth Edition. *Perceptual and Motor Skills*. Vol 74. 347-353.

Messick, S. (1989). "Validity." In R.L. Linn (Ed.), *Educational Measurement*, 3rd. Edition. New York: ACE/MacMillan.

Moss, P. A. (1994). "Can There Be Validity Without Reliability?" *Educational Researcher* 23(2), pp. 4-12.

Norusis, M. (1994). *SPSS Advanced Statistics 6.1*. Chicago: SPSS, Inc.

Norusis, M. (1994). *SPSS Base System Reference Guide 6.0*. Chicago: SPSS, Inc.

Norusis, M. (1994). *SPSS Base System User's Guide 6.0*. Chicago: SPSS Inc.

Norusis, M. (1994). *SPSS Professional Statistics 6.1*. Chicago: SPSS Inc.

Nugent, A. and G.J. Loabs. (1978). *Performance Test Objectivity: Comparison of Interrater Reliabilities of Three Observation Formats*. Paper presented at the Annual Meeting of the Military Testing Association. ERIC Document Reproduction Service (microfiche) number ED 186 474.

Schroeder, T.S. (1973). *Schroeder Composition Scale*. National Council of Teachers of English Committee on Research: The Research Instruments Project (TRIP), Urbana, IL: NCTE. ERIC Document Reproduction Service (microfiche) number ED 091760.

Shepard, L.A. (1993). "Evaluating Test Validity." In L. Darling-Hammond (Ed.) *Review of Research in Education 19*. Washington, DC: AERA, pp. 405-450.

Thompson, I. (1995). "A Study of Interrater Reliability of the ACTFL Oral Proficiency Interview in Five European Languages: Data from ESL, French, German, Russian, and Spanish." *Foreign Language Annals*. 28, No. 3: 407-422.

Umesh, U.N., R.A. Peterson and M.H. Sauber. (1989). "Interjudge agreement and the maximum value of Kappa." *Educational and Psychological Measurement*. 49:4, 835-850.

Appendix A. Examinee Instructions

Instructions for the Examinee

The Speaking Proficiency Test is a face-to-face test of your foreign language speaking ability. The test is administered by two testers and usually takes 15 to 45 minutes. The test is rated on a scale of 0 to 5. The testers will evaluate your ability to use the language appropriately when you participate in a conversation, obtain information from a native speaker, perform tasks, and speak at length.

The test is designed to assess your language proficiency in relation to that of an educated native speaker in a country where the language is spoken. You will not be tested on any specific professional specialty, nor on what you may have learned in a language course. In order to give you the opportunity to reach your highest level, the testers may at times use language more advanced than you feel you are able to handle.

Test Content:

1. **Conversation**
   Most of the test will be a conversation between you and the two testers. As with any conversation, a variety of topics will be covered.

2. **Situation**
   A tester will set up a role-playing situation that you and the tester will act out. You will not be asked to take the role of anyone except yourself.

3. **Information Gathering Task**
   You will be given the opportunity to interview one tester on a certain topic and then to report the information you learned to the other tester.

Hints for taking the test:

- Respond to questions or situations as fully as possible.

- If you are not comfortable with a topic for personal reasons, feel free to say so in a way that is natural within the conversation. However, if you use this privilege often, you may hurt your chance of demonstrating your true ability.

- Actively participate in the conversation. Feel free to ask questions, introduce topics, and ask for clarification when necessary.

Tester Script : Oral Summary by Testers of Important Points

To ensure that examinees know what to expect during the test, once the examinee enters the testing room, the testers read aloud the following information. Testers record the reading of these instructions along with the actual test.

Oral Summary of Instructions for the Examinee

(to be **read** by testers before beginning of test)

Have you had the opportunity to read the written test instruction sheet?

Do you have any questions about it?

REMINDERS

This is a proficiency test. We are trying to assess your language proficiency in relation to that of an educated native speaker of the language in which you are being tested.

Most of the test will be a conversation between you and the two testers, and it will last between 15-45 minutes.

We will cover a variety of topics. If you are uncomfortable with a particular topic, please let us know and we will go on to a different one. We are only interested in seeing how you handle the language.

A couple of activities other than conversation will be used. We will provide clear instructions for them later in the test.

Please feel free to take the initiative or ask for clarification at any time during the test.

Examinee Instructions: The Information Gathering Task

Your task is to elicit information and opinion from one of the testers in the test language on a topic which will be given to you. You will need to manage the interaction, to understand what you are told, and then to report in English (to the other tester) what you find out. If you do not understand something in the response to your question, ask for clarification or repetition. You may take notes. The tester will tell you what topic to address and when to give your report in English.

Appendix B. Pre-Test Questionnaire

Office Use Only Examinee ID#: Date: AM/PM

Unified Language Testing Plan
Speaking Proficiency Test
Pilot Study

Pre-Test Questionnaire

In order to help us validate this new speaking proficiency test, please take a moment to answer the following questions:

1) Sex: (Please circle the appropriate response.) Male Female

2) Present Age: ____________

3) Age when you began learning English: _______________

4) In what setting did you learn English? (Please circle all that apply.)

at home elementary school middle school high school
college in-country intensive language course
other: ____

5) Your native language(s): ____________________________________

6) Language(s) spoken in your home when you were a child: ____________________
____

7) Native language of others in your family (if different from your own): ____________
Relationship (spouse, children, etc.): ___________________________
Language you normally speak with him/her/them: ______________

8) How often do you:

a) speak in English?	☐ every day	☐ at least once a week	☐ at least once a month	☐ rarely	☐ never
b) listen to spoken English?	☐ every day	☐ at least once a week	☐ at least once a month	☐ rarely	☐ never
c) read in English?	☐ every day	☐ at least once a week	☐ at least once a month	☐ rarely	☐ never
d) write in English?	☐ every day	☐ at least once a week	☐ at least once a month	☐ rarely	☐ never

9) Foreign language learning & testing history:

Language learned include English	How long have you been learning this language?	When did you take your last proficiency test in this language?	Which agency administered the test? (LTD, DLI, FSI, Peace Corps, etc.)	What score did you receive? (speaking test only)
1.				
2.				
3.				

Appendix C.   Post-Test Questionnaires

Office use only:	Examinee ID#:	Date:	a.m./p.m.

Test number ________

Thank you for participating in this study. Please answer the following questions about your last test.

Circle one response for each question.

1. I felt that the quality of my English during the test was: a) *better than usual*
   b) *about average for me*
   c) *worse than usual*

   Why ?________________________________

2. I felt the testers heard a good sample of the English I know. a) *yes* b) *no*
3. The testers found the limits of my English ability. a) *yes* b) *no*
4. The test seemed a) *easy* b) *about right* c) *too hard*
5. I liked the **conversation** portion of the test. a) *yes* b) *no*
6. I felt this section tested a realistic use of language. a) *yes* b) *no*
7. I liked the topics we covered in this section. a) *yes* b) *no*

   Why or why not?________________________________

8. I liked the **situation** portion of the test. a) *yes* b) *no*
9. I felt the situation tested a realistic use of language. a) *yes* b) *no*
10. I liked the situation I was given. a) *yes* b) *no*

    Why or why not?________________________________

11. I liked the **information gathering task.** a) *yes* b) *no*

    Why or why not?________________________________

12. I felt this task tested a realistic use of language. a) *yes* b) *no*

Please write any additional comments on the back of this page.

Office use only	Examinee ID#	Date	a.m./p.m.

Summary (to be completed after the fourth test)

In your own opinion

1. **Rank your four tests from easiest to hardest.**

Fill in test number.

Easiest ______ 2nd easiest ______ 3rd easiest ______ Hardest ______

What are the main reasons for this ranking?

____

____

2. **Rank your four tests according to the quality of your language performance.**

Fill in test number.

Best ______ 2nd best ______ 3rd best ______ Worst ______

What are the main reasons for this ranking?

____

____

General comments on the four tests:

Appendix D. Frequency Charts

88

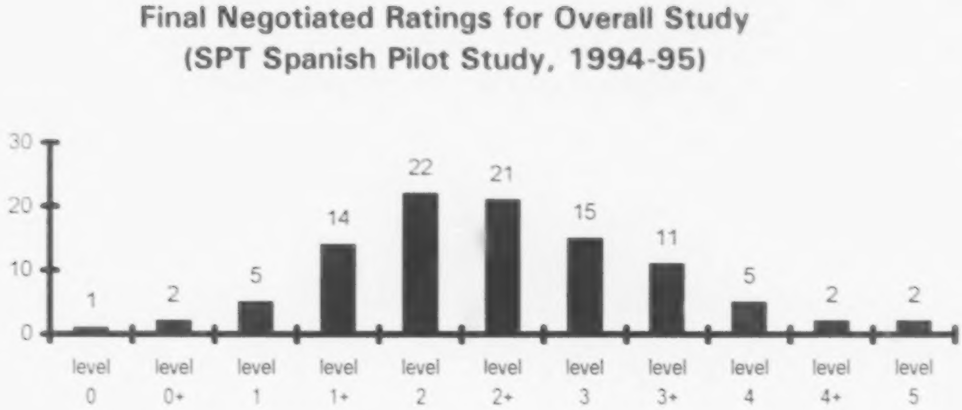


Chart D-1. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for the overall Spanish study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	10.000
Skewedness	0.2168
Kurtosis	0.0805
K-S Lilliefors test results	stat 0.1741 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

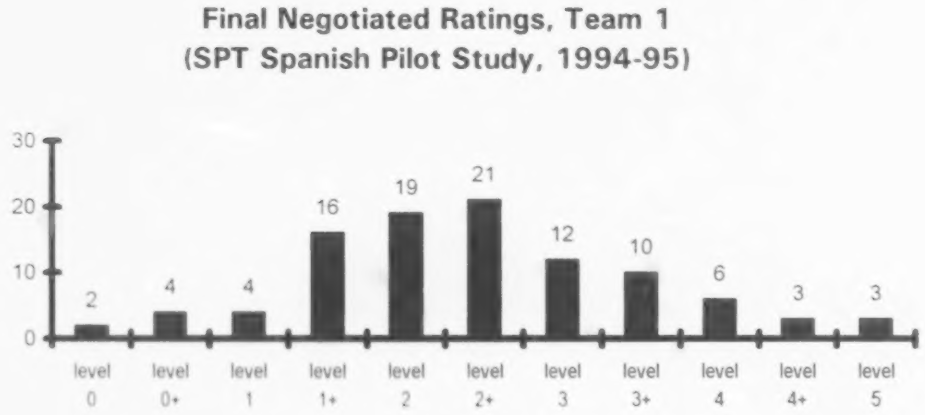


Chart D-2. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for team 1 for the overall Spanish study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	12.000
Skewedness	0.1893
Kurtosis	0.0213
K-S Lilliefors test results	stat 0.1651 p .0000**

One-tailed probability value (p) is reported.
*α = .05; *p < .05; **p < .01*

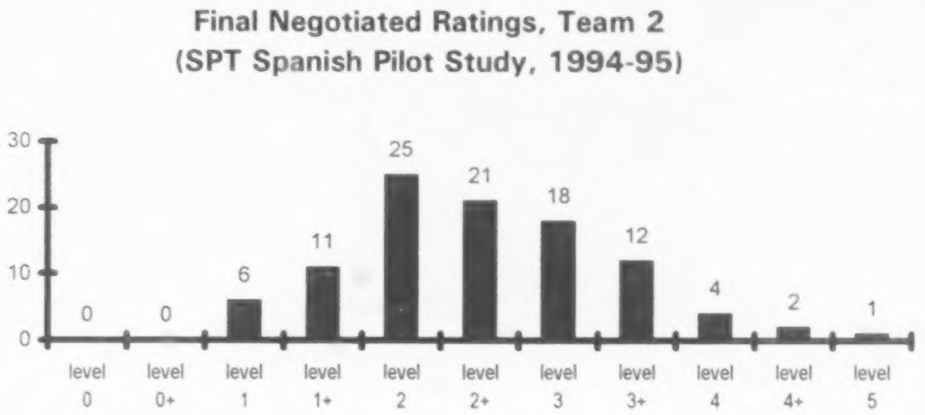


Chart D-3. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for team 2 for the overall Spanish study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	10.000
Skewedness	0.3357
Kurtosis	-0.1204
K-S Lilliefors test results	stat 0.1903 p .0000**

One-tailed probability value (p) is reported.
*α = .05; *p< .05; **p< .01*

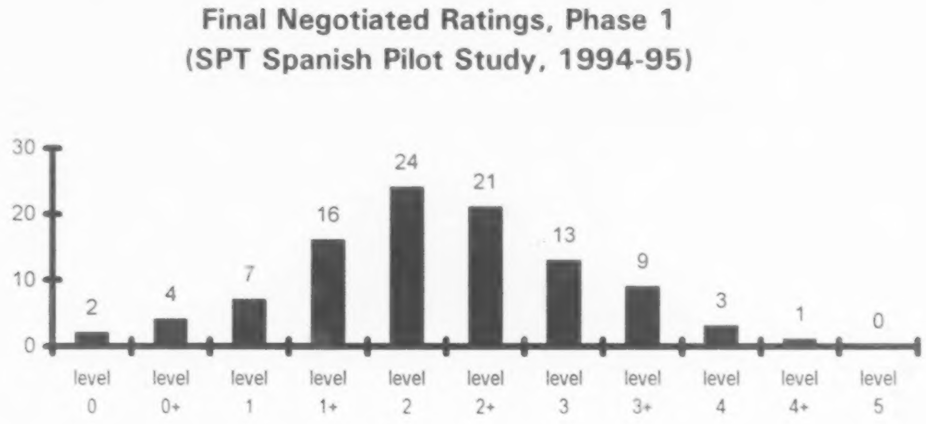


Chart D-4. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for phase 1 of the Spanish study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2
Interquartile Range	12.000
Skewedness	-0.0380
Kurtosis	0.1114
K-S Lilliefors test results	stat 0.1769 p .0000**

One-tailed probability value (p) is reported.
*α = .05; *p< .05; **p< .01*

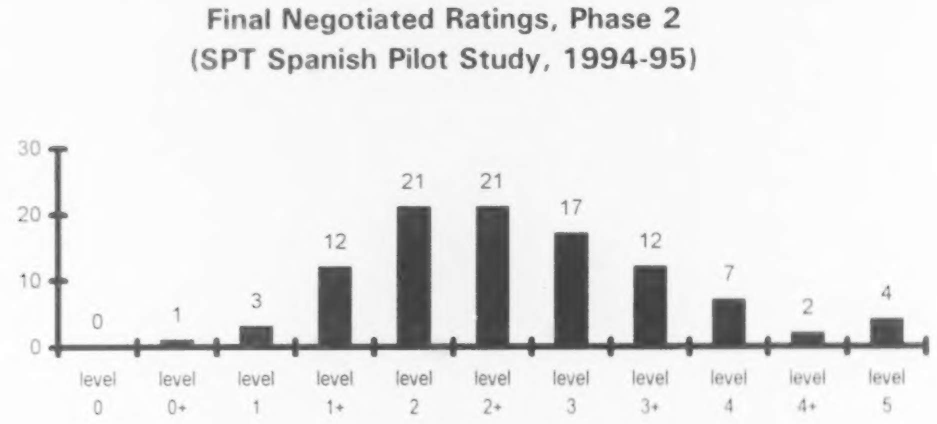


Chart D-5. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for phase 2 of the Spanish study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	16.000
Skewedness	0.4322
Kurtosis	-0.2216
K-S Lilliefors test results	stat 0.1719 p .0000**

One-tailed probability value (p) is reported.
*α = .05; *p< .05; **p< .01*

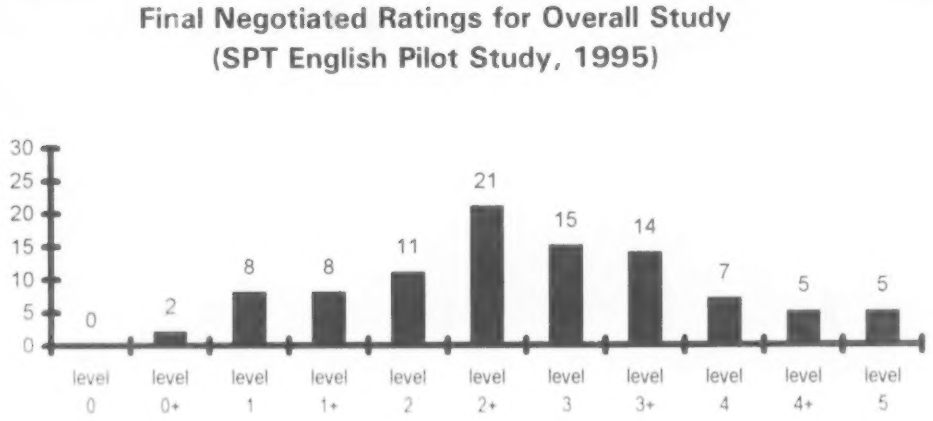


Chart D-6. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for the overall English study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data do not seem to be distributed normally; that is, the data do not fit under a classical bell-shaped curve.

Normality Data	
Median	3
Interquartile Range	18.000
Skewedness	2.0847
Kurtosis	6.2847
K-S Lilliefors test results	stat 0.1932 p .0000**

One-tailed probability value (p) is reported.
*α = .05; *p< .05; **p< .01*

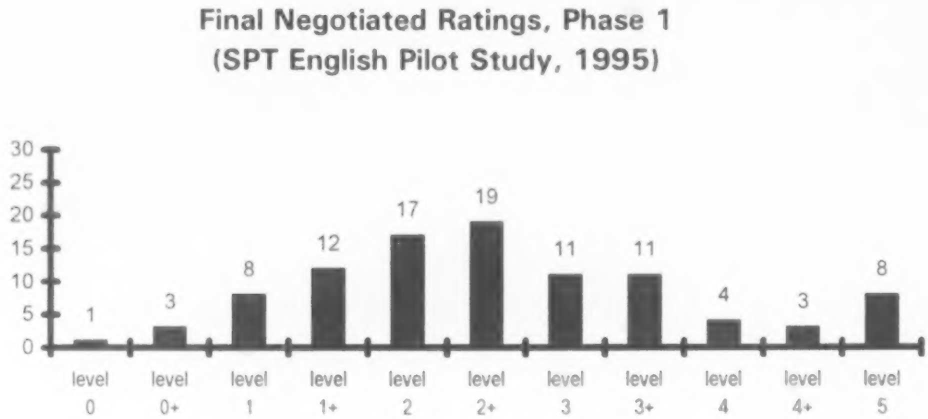

Final Negotiated Ratings, Phase 1
(SPT English Pilot Study, 1995)

Chart D-7. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for phase 1 of the English study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data do not seem to be distributed normally; that is, the data do not fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	18.000
Skewedness	2.0767
Kurtosis	6.5426
K-S Lilliefors test results	stat 0.1995 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

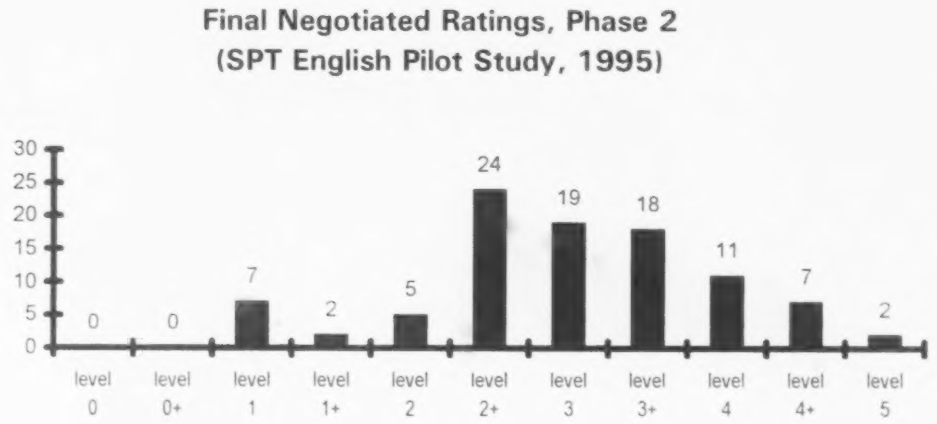

Final Negotiated Ratings, Phase 2
(SPT English Pilot Study, 1995)

Chart D-8. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for phase 2 of the English study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data do not seem to be distributed normally; that is, the data do not fit under a classical bell-shaped curve.

Normality Data	
Median	3
Interquartile Range	11.500
Skewedness	2.2557
Kurtosis	6.7171
K-S Lilliefors test results	stat 0.2449 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

Final Negotiated Ratings, Testing Pair 1
(SPT English Pilot Study, 1995)

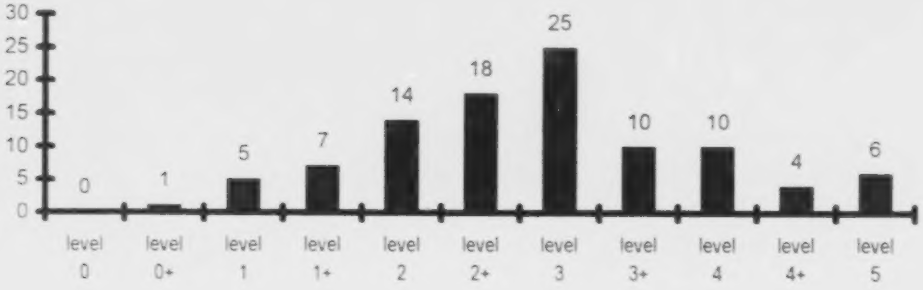


Chart D-9. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for testing pair 1 in the overall English study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data do not seem to be distributed normally; that is, the data do not fit under a classical bell-shaped curve.

Normality Data		
Median	3	
Interquartile Range	18.000	
Skewedness	2.3373	
Kurtosis	7.6155	
K-S Lilliefors test results	stat p	0.2358 .0000**

One-tailed probability value (p) is reported.
*α = .05;   *p< .05; **p< .01*

97

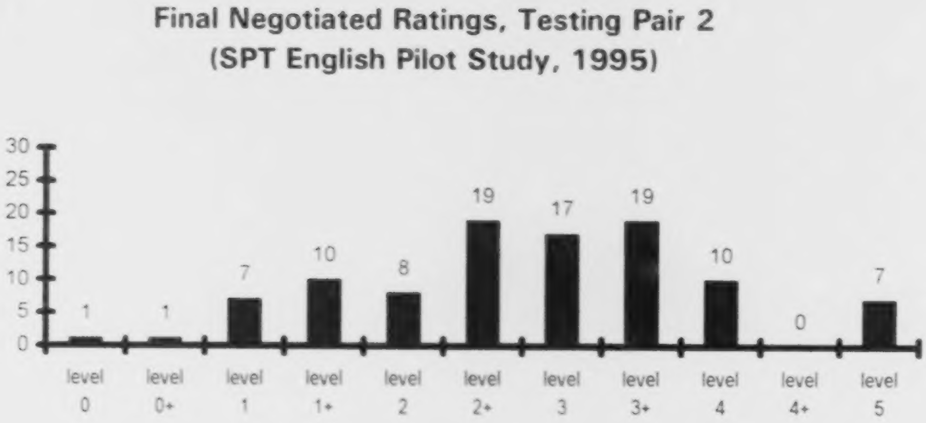

Final Negotiated Ratings, Testing Pair 2
(SPT English Pilot Study, 1995)

Chart D-10. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for testing pair 2 in the overall English study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data do not seem to be distributed normally; that is, the data do not fit under a classical bell-shaped curve.

Normality Data	
Median	3
Interquartile Range	18.000
Skewedness	2.0707
Kurtosis	6.8572
K-S Lilliefors test results	stat 0.2209 p .0000**

One-tailed probability value (p) is reported.
*α = .05; *p< .05; **p< .01*

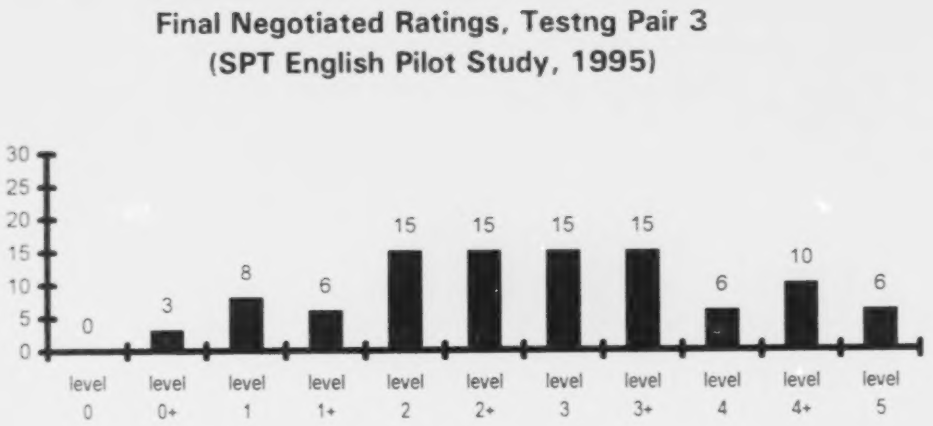

Final Negotiated Ratings, Testng Pair 3
(SPT English Pilot Study, 1995)

Chart D-11. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for testing pair 3 in the overall English study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data do not seem to be distributed normally, that is, the data do not fit under a classical bell-shaped curve.

Normality Data	
Median	3
Interquartile Range	20.000
Skewedness	1.8819
Kurtosis	4.7008
K-S Lilliefors test results	stat 0.1721 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

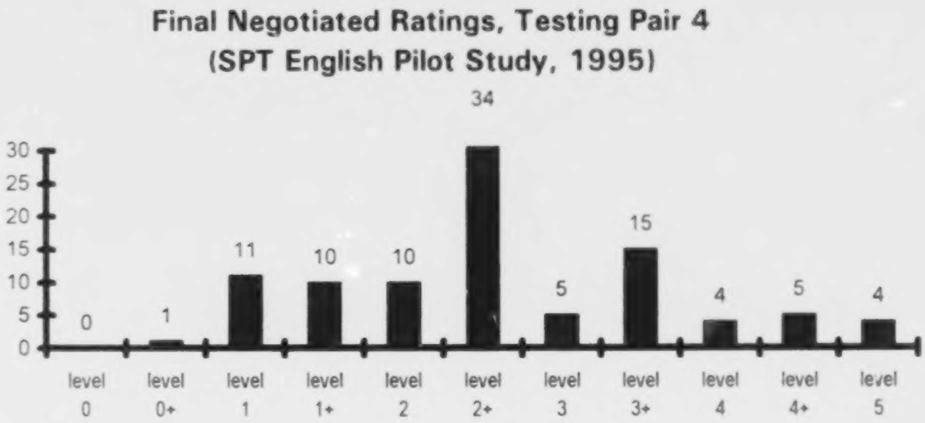

Final Negotiated Ratings, Testing Pair 4
(SPT English Pilot Study, 1995)

Chart D-12. The data in this chart reflect the distribution of examinees' final negotiated ratings (live ratings only) across the ILR scale for testing pair 4 in the overall English study. The table below contains data related to the distribution of the scores across the ILR levels. In general, these data do not seem to be distributed normally; that is, the data do not fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	18.000
Skewedness	2.1666
Kurtosis	8.1063
K-S Lilliefors test results	stat 0.1940 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p < .05; **p < .01*	

Appendix E. Summary Spanish Results

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Overall Study

Table E-1. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square
Spanish Pilot Study: Overall Study

	FSI			FBI			DLI		
Agency	x^2	df	2-tailed p	x^2	df	2-tailed p	x^2	df	2-tailed p
CIA	399.424	90	.0000**	433.047	100	.0000**	346.143	100	.0000**
DLI	442.534	100	.0000**	344.284	100	.0000**			
FBI	393.586	100	.0000**						

The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by each agency for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table E-2. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test
Spanish Pilot Study: Overall Study

Agency	**Median**	**Interquartile Range (IQR)**
CIA	2+	18.00
DLI	2+	12.00
FBI	2+	10.00
FSI	2	12.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
58.0280	3	.0000**

The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by each agency for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table E-3. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Spanish Pilot Study: Overall Study

	FSI		FBI		DLI	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -6.4512 p = .0000**	z = -6.7416 p = .0000**	z = -3.1845 p = .0014**	z = -3.4857 p = .0003**	z = -4.3994 p = .0000**	z = -4.5957 p = .0000**
DLI	z = -4.2420 p = .0000**	z = -3.8158 p = .0000**	z = -.8093 p = .4179	z = -1.3148 p = .1926		
FBI	z = -4.7976 p = .0000**	z = -5.0468 p = .0000**				

*The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by each agency for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α = .05. *p < .05. **p < .01*

Table E-4. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula Spanish Pilot Study: Overall Study

Agency	**FSI**	**FBI**	**DLI**
CIA	758	737	791
DLI	799	773	
FBI	740		

The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by each agency for live ratings only.

Table E-5. Interagency Reliability as Measured by Percent Level of Agreement Spanish Pilot Study: Overall Study

Agency	**FSI**	**FBI**	**DLI**
CIA	3783	4285	4579
DLI	4453	3750	
FBI	4310		

The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by each agency for live ratings only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Team 1 Only*

*Each of the two teams was made up of a pair of testers from each participating agency

Team 1	Team 2
Pair 1 CIA	*Pair 1 CIA*
Pair 2 DLI	*Pair 2 DLI*
Pair 3 FBI	*Pair 3 FBI*
Pair 4 FSI	*Pair 4 FSI*

Table E-6. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square Spanish Pilot Study Team 1 Only

Agency	**FSI**			**FBI**			**DLI**		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	224.862	90	.0000**	256.637	100	.0000**	236.027	100	.0000**
DLI	236.391	90	.0000**	222.305	100	.0000**			
FBI	216.276	90	.0000**						

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 1 testing pair from each agency for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported.

α .05; *p .05; **p .01

Table E-7. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test Spanish Pilot Study Team 1 Only

Agency	**Median**	**Interquartile Range (IQR)**
CIA	2+	18.00
DLI	2+	12.00
FBI	2+	10.00
FSI	2	12.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
29.3689	3	.0000**

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 1 testing pair from each agency for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

α .05; *p .05; **p .01

Table E-8. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign tests, Spanish Pilot Study: Team 1 Only

	FSI		FBI		DLI	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -4.2791 p = .0000**	z = -4.1667 p = .0000**	z = -1.1070 p = .2679	z = -1.4856 p = .1308	z =-3.0197 p = .0015**	z = -3.4816 p = .0002**
DLI	z = -2.4174 p = .0147**	z = -1.8865 p = .0579	z = -2.6924 p = .0062**	z = -2.6879 p = .0068**		
FBI	z = -4.0528 p = .0000**	z = -3.8431 p = .0001**				

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 1 testing pair from each agency for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

*α .05; *p .05; **p .01*

Table E-9. Interagency Reliability as Measured by Kendall Correlation Formula Tau-b Spanish Pilot Study: Team 1 Only

Agency	**FSI**	**FBI**	**DLI**
CIA	.752	.780	.794
DLI	.801	.790	
FBI	.750		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 1 testing pair from each agency for live ratings only.

Table E-10 Interagency Reliability as Measured by Percent Level of Agreement Spanish Pilot Study: Team 1 Only

Agency	**FSI**	**FBI**	**DLI**
CIA	.3793	.4629	.4310
DLI	.4603	.3220	
FBI	.3389		

The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by the team 1 testing pair for live ratings only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Team 2 Only*

*Each of the two teams was made up of a pair of testers from each participating agency.

Team 1		*Team 2*	
Pair 1	*CIA*	*Pair 1*	*CIA*
Pair 2	*DLI*	*Pair 2*	*DLI*
Pair 3	*FBI*	*Pair 3*	*FBI*
Pair 4	*FSI*	*Pair 4*	*FSI*

Table E-11. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square Spanish Pilot Study: Team 2 Only

Agency	**FSI**			**FBI**			**DLI**		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	98.401	35	.0000**	91.915	35	.0000**	107.945	49	.0000**
DLI	161.459	48	.0000**	120.498	48	.0000**			
FBI	102.910	36	.0000**						

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 2 testing pair from each agency for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported.

$\alpha = .05$; $*p < .05$; $**p < .01$

Table E-12. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test Spanish Pilot Study: Team 2 Only

Agency	**Median**	**Interquartile Range (IQR)**
CIA	2+	18.00
DLI	2+	10.00
FBI	2+	10.00
FSI	2	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
32.75	3	.0000**

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 2 testing pair from each agency for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

$\alpha = .05$; $*p < .05$; $**p < .01$

Table E-13. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Spanish Pilot Study: Team 2 Only

	FSI		FBI		DLI	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -4.8251 p = .0000**	z = -5.2223 p = .0000**	z = -3.2858 p = .0004**	z = -3.2329 p = .0008**	z = -3.2126 p = .0012**	exact p = .0041**
DLI	z = -3.5597 p = .0000**	z = -3.3588 p = .0006**	z = -1.5549 p = .1268	z = -.9129 p = .3648		
FBI	z = -2.6051 p = .0076**	z = -3.0792 p = .0014**				

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 2 testing pair from each agency for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, these values are reported as ***exact****. Two-tailed probability value (p) is reported. α = .05; *p < .05; **p < .01*

Table E-14. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula Spanish Pilot Study: Team 2 Only

Agency	**FSI**	**FBI**	**DLI**
CIA	.783	.716	.805
DLI	.790	.795	
FBI	.756		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 2 testing pair from each agency for live ratings only.

Table E-15. Interagency Reliability as Measured by Percent Level of Agreement Spanish Pilot Study: Team 2 Only

Agency	**FSI**	**FBI**	**DLI**
CIA	.4150	.3921	.4897
DLI	.4285	.4339	
FBI	.5263		

The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by the team 2 testing pair for live ratings only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Phase 1 Only

Table E-16. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square Spanish Pilot Study: Phase 1 Only

Agency	FSI			FBI			DLI		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	192.648	64	.0000**	176.350	56	.0000**	155.106	72	.0000**
DLI	168.067	72	.0000**	160.491	72	.0000**			
FBI	178.669	64	.0000**						

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by both testing pairs from each agency during the first half of the study for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table E-17. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test Spanish Pilot Study: Phase 1 Only

Agency	Median	Interquartile Range (IQR)
CIA	2	12.00
DLI	2+	11.00
FBI	2+	10.00
FSI	2	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
15.2634	3	.0016**

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by both testing pairs from each agency during the first half of the study for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table E-18. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Spanish Pilot Study: Phase 1 Only

	FSI		**FBI**		**DLI**	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -3.4800 p = .0001**	z = -3.7262 p = .0001**	z = -.8061 p = .4370	exact p = .4049	z = -1.7669 p = .0802	exact p = .0433**
DLI	z = -2.6080 p = .0080**	z = -1.8570 p = .0572	z = -.9519 p = .3547	z = -1.3229 p = .1834		
FBI	z = -3.2070 p = .0012**	z = -3.1038 p = .0007**				

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by both testing pairs from each agency during the first half of the study for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it is possible to calculate exact probability values, these values are reported as **exact**. *Two-tailed probability value (p) is reported. α = .05; *p < .05; **p < .01*

Table E- 19. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula Spanish Pilot Study: Phase 1 Only

Agency	**FSI**	**FBI**	**DLI**
CIA	.797	.747	.809
DLI	.791	.790	
FBI	.783		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by both testing pairs from each agency during the first half of the study for live ratings only.

Table E-20. Interagency Reliability as Measured by Percent Level of Agreement Spanish Pilot Study: Phase 1 Only

Agency	**FSI**	**FBI**	**DLI**
CIA	.4800	.5000	.4680
DLI	.4528	.4375	
FBI	.4117		

The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by both testing pairs from each agency during the first half of the study for live ratings only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Phase 2 Only

Table E-21. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square Spanish Pilot Study: Phase 2 Only

Agency	FSI			FBI			DLI		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	179.744	64	.0000**	152.551	64	.0000**	186.089	64	.0000**
DLI	238.611	72	.0000**	180.793	64	.0000**			
FBI	199.991	64	.0000**						

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by both testing pairs from each agency during the second half of the study for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01

Table E-22. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test Spanish Pilot Study: Phase 2 Only

Agency	Median	Interquartile Range (IQR)
CIA	3	20.00
DLI	2+	18.00
FBI	2+	14.00
FSI	2+	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
45.45	3	.0000**

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by both testing pairs from each agency during the second half of the study for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01

Table E-23. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign test, Spanish Pilot Study: Phase 2 Only

	FSI		FBI		DLI	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -5.4111 p = .0000**	z = -5.4899 p = .0000**	z = -3.4184 p = .0005**	z = -3.5000 p = .0003**	z = -4.3104 p = .0000**	z = -4.4194 p = .0000**
DLI	z = -3.3280 p = .0008**	z = -3.2880 p = .0006**	z = -.5760 p = .5768	z = -.6247 p = .5304		
FBI	z = -3.5184 p = .0003**	z = -3.8333 p = .0001**				

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned by both testing pairs from each agency during the second half of the study for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01*

Table E-24. Interagency Reliability as Measured by Kendall Tau-B Correlation Formula Spanish Pilot Study: Phase 2 Only

Agency	**FSI**	**FBI**	**DLI**
CIA	.722	.721	.771
DLI	.798	.745	
FBI	.701		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by both testing pairs from each agency during the second half of the study for live ratings only.

Table E-25. Interagency Reliability as Measured by Percent Level of Agreement Spanish Pilot Study: Phase 2 Only

Agency	**FSI**	**FBI**	**DLI**
CIA	.2950	.3793	.4576
DLI	.4393	.3492	
FBI	.4461		

The data in this table are the results of comparing all of the Spanish final negotiated ratings assigned by both testing pairs from each agency during the second half of the study for live ratings only.

Inter-Rater Reliability
Summary Results: Non-Parametric Correlations Between Individual Final Ratings Within Testing Pairs
Spanish Pilot Study: Overall Study

Table E-26. Inter-Rater Reliability as Measured by Kendall Tau-b Correlation Formula
Spanish Pilot Study: Overall Study

CIA	DLI	FBI	FSI
n=125	n=125	n=125	n=125
.929	.958	.893	.998

The data in this table are the results of comparing the Spanish individual tester ratings assigned by the testers in each pair from each agency for live ratings only.

Table E-27. Inter-Rater Reliability as Measured by Kendall Tau-b Correlation Formula
Spanish Pilot Study: Team 1 Only

CIA	DLI	FBI	FSI
n=63	n=63	n=63	n=63
.904	1.000	.864	1.000

The data in this table are the results of comparing the Spanish individual tester ratings assigned by the testers in the team 1 pair from each agency for live ratings only.

Table E-28. Inter-Rater Reliability as Measured by Kendall Tau-b Correlation Formula
Spanish Pilot Study: Team 2 Only

CIA	DLI	FBI	FSI
n=62	n=62	n=62	n=62
.960	.916	.936	.995

The data in this table are the results of comparing the Spanish individual tester ratings assigned by the testers in the team 2 pair from each agency for live ratings only.

Table E-29. Inter-Rater Reliability as Measured by Kendall Tau-b Correlation Formula
Spanish Pilot Study: Phase 1 Only

CIA	DLI	FBI	FSI
n=56	n=56	n=56	n=56
.899	.951	.841	1.000

The data in this table are the results of comparing the Spanish individual tester ratings assigned by the testers in each pair from each agency for the first half of the study for live ratings only.

Table E-30. Inter-Rater Reliability as Measured by Kendall Tau-b Correlation Formula Spanish Pilot Study: Phase 2 Only

CIA	DLI	FBI	FSI
n=68	n=68	n=68	n=68
.945	.967	.939	1.000

The data in this table are the results of comparing the Spanish individual tester ratings assigned by the testers in each pair from each agency for the second half of the study for live ratings only.

Table E-31. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Kendall Tau-b Correlation Formula Taped Ratings Only, Spanish Pilot Study: Overall Study

CIA	DLI	FBI	FSI
n=48	n=48	n=48	n=48
.908	.964	845	1.000

The data in this table are the results of comparing the Spanish individual tester ratings for taped ratings only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Overall Study

Table E-32. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square
Spanish Pilot Study: Overall Study

Test Order	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	301.531	100	.0000**	372.148	90	.0000**	355.733	100	.0000**
Second	414.705	100	.0000**	407.014	100	.0000**			
Third	379.894	100	.0000**						

The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

α .05; *p .05; **p 01

Table E-33. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test
Spanish Pilot Study: Overall Study

Test Order	Median	Interquartile Range (IQR)
First	2+	10.00
Second	2+	10.00
Third	2+	10.00
Fourth	2+	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
12.1542	3	0056**

The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α 05; *p .05; **p 01

Table E-34. Test Order Effects as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Spanish Pilot Study: Overall Study

	Fourth		**Third**		**Second**	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
First	z = -2.8767 p = .0032**	z = -4.2172 p = .0000**	z = -2.3215 p = .0222*	z = -1.7638 p = .0728	z = -1.4543 p = .1459	z = -1.6002 p = .1050
Second	z = -2.0825 p = .0357*	z = -1.3970 p = .1584	z = -0.8299 p = .4078	z = -1.3148 p = .1838		
Third	z = -0.3119 p = .7542	z = 0.0000 p = 1.0000				

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α = .05; *p < .05; **p < .01*

Table E-35. Test Order Effects as Measured by Kendall Tau-b Correlation Formula Spanish Pilot Study: Overall Study

Test Order	**Fourth**	**Third**	**Second**
First	.715	.733	.751
Second	.734	.706	
Third	.704		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Team 1 Only

Table E-36. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square
Spanish Pilot Study: Team 1 Only

Test Order	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	191.162	100	.0001**	198.517	90	.0000**	202.744	100	.0000**
Second	286.691	100	.0000**	228.318	90	.0000**			
Third	221.855	90	.0000**						

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 1 testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01.*

Table E-37. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test
Spanish Pilot Study: Team 1 Only

Test Order	Median	Interquartile Range (IQR)
First	2+	12.00
Second	2+	10.00
Third	2+	12.00
Fourth	2+	12.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
4.4098	3	.2200

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 1 testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01*

Table E-38. Test Order Effects as Measured by Kendall Tau-b Correlation Formula
Spanish Pilot Study: Team 1 Only

Test Order	Fourth	Third	Second
First	.696	.766	.720
Second	.774	.708	
Third	.737		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 1 testers in the tests administered first, second, third and fourth to each examinee for live ratings only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Team 2 Only

Table E-39. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square Spanish Pilot Study: Team 2 Only

Test Order	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	113.060	48	.0001**	118.506	42	.0000**	150.426	49	.0000**
Second	101.860	56	.0276*	129.396	56	.0000**			
Third	109.675	49	.0000**						

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 2 testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01

Table E-40. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test Spanish Pilot Study: Team 2 Only

Test Order	Median	Interquartile Range (IQR)
First	2+	10.00
Second	2+	10.00
Third	2+	10.00
Fourth	2+	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
10.6848	3	.0122*

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 2 testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01

Table E-41. Test Order Effects as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Spanish Pilot Study: Team 2 Only

	Fourth		Third		Second	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
First	z = -3.0569 p = .0007**	z = -3.7262 p = .0002**	z = -1.7962 p = .0801	z = -1.7650 p = .0756	z = -0.9553 p = .3279	z= -0.0000 p = 1.0000
Second	z = -2.0268 p = .0412*	z = -1.6440 p = .0940	z = -1.0699 p = .2987	z = -1.2374 p = .2135		
Third	z = -0.3000 p = .7721	z = -0.1890 p = .8467				

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 2 testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table E-42. Test Order Effects as Measured by Kendall Tau-b Correlation Formula Spanish Pilot Study: Team 2 Only

Test Order	Fourth	Third	Second
First	.753	.686	.810
Second	.703	.714	
Third	.656		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by the team 2 testers in the tests administered first, second, third, and fourth to each examinee for live ratings only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Phase 1

Table E-43. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square
Spanish Pilot Study: Phase 1 Only

Test Order	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	176.894	72	.0000**	159.232	56	.0000**	156.591	72	.0000**
Second	181.587	81	.0000**	170.128	63	.0000**			
Third	155.209	56	.0000**						

The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only during the first half of the study. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table E-44. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test
Spanish Pilot Study: Phase 1 Only

Test Order	Median	Interquartile Range (IQR)
First	2	10.00
Second	2	10.00
Third	2+	12.00
Fourth	2+	12.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
7.1908	3	.0651

The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only during the first half of the study. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table E-45. Test Order Effects as Measured by Kendall Tau-b Correlation Formula
Spanish Pilot Study: Phase 1 Only

Test Order	Fourth	Third	Second
First	.804	.774	.756
Second	.774	.745	
Third	.786		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only during the first half of the study.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Phase 2 Only

Table E-46. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square Spanish Pilot Study: Phase 2 Only

Test Order	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	164.471	72	.0000**	185.081	72	.0000**	193.074	72	.0000**
Second	159.977	56	.0000**	187.884	72	.0001**			
Third	179.219	72	.0000**						

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only during the second half of the study. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01*

Table E-47. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test Spanish Pilot Study: Phase 2 Only

Test Order	Median	Interquartile Range (IQR)
First	2+	10.00
Second	2+	18.00
Third	2+	18.00
Fourth	2+	18.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
6.6868	3	.0798

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only during the second half of the study. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01*

Table E-48. Test Order Effects as Measured by Kendall Tau-b Correlation Formula Spanish Pilot Study: Phase 2 Only

Test Order	Fourth	Third	Second
First	.643	.718	.739
Second	.689	.674	
Third	.640		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only during the second half of the study.

BEST COPY AVAILABLE

Effects on Reliability Caused by Time of Administration
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Overall Study

Table E-49. Test Slot Effects as Measured by Non-Parametric Pearson Chi-Square
Spanish Pilot Study: Overall Study

Test Slot	**2:30 p.m.**			**1:00 p.m.**			**10:30 a.m.**		
	x^2	df	p	x^2	df	p	x^2	df	p
9:00 a.m.	176.343	72	.0002**	241.736	72	.0000**	222.814	72	.0000**
10:30 a.m.	242.046	72	.0000**	311.629	81	.0000**			
1:00 p.m.	264.003	81	.0000**						
Morning tests compared to afternoon tests							449.509	80	.0000**

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned by all testing pairs in the tests administered in each testing slot for each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01*

Table E-50. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test
Spanish Pilot Study: Overall Study

Test Slot	**Median**	**Interquartile Range (IQR)**
9:00 a.m.	2+	10.00
10:30 a.m.	2+	10.00
1:00 p.m.	2+	10.00
2:30 p.m.	2+	10.00
Friedman Two-way Anova Chi-Square of Ranks (comparing all four testing slots)		
x^2	df	2-tailed p value
1.1135	3	.7803
Friedman Two-way Anova Chi-Square of Ranks (comparing all a.m. tests to all p.m. tests)		
x^2	df	2-tailed p value
.2747	1	.6792

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned by all testing pairs in the tests administered in each testing slot for each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01*

Table E-51. Test Slot Effects as Measured by Kendall Tau-b Correlation Formula
Spanish Pilot Study: Overall Study

Test Order	Fourth	Third	Second
First	.658	.738	.759
Second	.695	.722	
Third	.706		
Morning Tests Compared to Afternoon Tests			.708

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by all testing pairs in the tests administered in each testing slot for each examinee for live ratings only.

Interagency Reliability for Taped Ratings Only (5-8)
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Overall Study

Table E-52. Interagency Reliability for Taped Ratings as Measured by Non-Parametric Pearson Chi-Square, Spanish Pilot Study: Overall Study

Agency	FSI			FBI			DLI		
	x^2	df	2-tailed p	x^2	df	2-tailed p	x^2	df	2-tailed p
CIA	88.468	42	0009**	81.908	42	.0047**	111.367	42	.0000**
DLI	73.876	36	0000**	109.980	42	.0000**			
FBI	73.077	42	0050**						

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned by all testing pairs from each agency for taped ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. α .05; *p .05; **p .01*

Table E-53. Interagency Reliability for Taped Ratings as Measured by Friedman Chi-Square of Ranks Test, Spanish Pilot Study: Overall Study

Agency	Median	Interquartile Range (IQR)
CIA	2+	10.00
DLI	2+	12.00
FBI	2+	10.00
FSI	2	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
14.2524	3	0015**

*The data in this table are the results of comparing the Spanish final negotiated ratings assigned by all testing pairs from each agency for taped ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. wo-tailed probability value (p) is reported. α .05; *p .05; **p .01*

Table E-54. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Spanish Pilot Study: Overall

	FSI		**FBI**		**DLI**	
Test	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -3.0364 p = .0017**	z = -3.1038 p = .0018**	z = -0.5614 p = .6087	z = -0.3714 p = .7098	z = -1.2779 p = .1989	exact p = .1078**
DLI	z = -1.4540 p = .1556	z = -1.6432 p = .0964	z = -1.5753 p = .1246	exact p = .1078		
FBI	z = -2.3461 p = .0188*	z = -2.5725 p = .0097**				

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by all testing pairs from each agency for taped ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, these values are reported as ***exact****. Two-tailed probability value (p) is reported.* α = *.05;* **p < .05, **p < .01*

Table E-55. Interagency Reliability for Taped Ratings as Measured by Kendall Tau-b Correlation Formula, Spanish Pilot Study: Overall Study

Agency	**FSI**	**FBI**	**DLI**
CIA	.676	.664	.748
DLI	.708	.814	
FBI	.630		

The data in this table are the results of comparing the Spanish final negotiated ratings assigned by all testing pairs from each agency for taped ratings only.

Intra-Agency Reliability
Comparison of Live and Taped Ratings (1/5.2/6.3/7.4/8)
Summary Results: Non-Parametric Analyses of Variance
Spanish Pilot Study: Overall Study

Table E-56. Intra-Agency Reliability Between Live and Taped Ratings as Measured by Non-Parametric Pearson Chi-Square, Spanish Pilot Study: Overall Study

Agency	x^2	df	2-tailed p
CIA	68.669	30	.0000**
DLI	94.532	48	.0000**
FBI	82.828	42	.0115 *
FSI	87.290	42	.0004**

The data in this table are the results of comparing the Spanish taped final negotiated ratings with their respective live final negotiated ratings. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $^{*}p < .05$; $^{**}p < .01$

Table E-57. Intra-Agency Reliability Between Live and Taped Ratings as Measured by Friedman Chi-Square of Ranks Test, Spanish Pilot Study: Overall Study

	Median	Interquartile Range (IQR)
Live Rating--CIA	2+	18.00
Live Rating--DLI	2+	12.00
Live Rating--FBI	2+	10.00
Live Rating--FSI	2	10.00
Taped Rating--CIA	2+	10.00
Taped Rating--DLI	2+	12.00
Taped Rating--FBI	2+	10.00
Taped Rating--FSI	2	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	df	2-tailed p value
41.5901	7	.0000**

The data in this table are the results of comparing the Spanish taped final negotiated ratings with their respective live final negotiated ratings. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $^{*}p < .05$; $^{**}p < .01$

Table E-58. Intra-Agency Reliability Between Live and Taped Ratings as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Spanish Pilot Study: Overall

	Wilcoxon	Sign
CIA	z = -2.5949 p = .0090**	exact p = .0118*
DLI	z = -1.5944 p = .1197	z = -0.9806 p = .3230
FBI	z = -0.5451 p = .5908	z = -0.1890 p = .8482
FSI	z = -0.5433 p = .5900	z = -0.7698 p = .4353

The data in this table are the results of comparing the Spanish taped final negotiated ratings with their respective live negotiated ratings. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values these values are reported as ***exact****. Two-tailed probability value (p) is reported. α = .05; *p < .05; **p < .01*

Table E-59. Intra-Agency Reliability Between Live and Taped Ratings as Measured by Kendall Tau-b Correlation Formula, Spanish Pilot Study: Overall Study

CIA	DLI	FBI	FSI
.770	.744	.652	.770

The data in this table are the results of comparing the Spanish taped final negotiated ratings with their respective live final negotiated ratings.

Appendix F. Summary English Results

Inter-Pair Reliability
Summary Results: Non-Parametric Analyses of Variance
English Pilot Study: Overall Study

Table F-1. Inter-Pair Reliability as Measured by Non-Parametric Pearson Chi-Square, English Pilot Study: Overall Study

	Pair 4			**Pair 3**			**Pair 2**		
(n=68)	x^2	df	2-tailed p	x^2	df	2-tailed p	x^2	df	2-tailed p
Pair 1	330.344	81	.0000**	235.981	81	.0000**	272.373	81	.0000**
Pair 2	275.948	81	.0000**	248.175	81	.0000**			
Pair 3	243.423	81	.0000**						

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. $\alpha = .05$; *p<.05; **p<.01

Table F-2. Inter-Pair Reliability as Measured by Friedman Chi-Square of Ranks test English Pilot Study: Overall Study

	Median	**Interquartile Range (IQR)**
Pair 1	3	18.00
Pair 2	3	18.00
Pair 3	3	20.00
Pair 4	2+	18.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
22.104	3	.0000**

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. $\alpha = .05$; *p<.05; **p<.01

Table F-3. Inter-Pair Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, English Pilot Study: Overall Study

	Pair 4		Pair 3		Pair 2	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
Pair 1	z =-3.5703 p = .0004**	z = -4.3729 p = .0000**	z = -0.5201 p = .6097	z = -0.9864 p = .3260	z =-0.4997 p = .6167	z = -0.6576 p = .5122
Pair 2	z = -2.5815 p = .0094**	z = -3.0822 p = .0014**	z = -0.0074 p = 1.0000	z = -0.4867 p = .6293		
Pair 3	z = -2.4858 p = .0099**	z = -2.7045 p = .0063**				

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

*α = .05; *p<.05; **p<.01*

Table F-4. Inter-Pair Reliability as Measured by Kendall Tau-b Correlation Formula English Pilot Study: Overall Study

	Pair 4	Pair 3	Pair 2
Pair 1	.853	.798	.827
Pair 2	.836	.798	
Pair 3	.812		

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings only.

Table F-5. Inter-Pair Reliability as Measured by Percent Level of Agreement English Pilot Study: Overall Study

	Pair 4	Pair 3	Pair 2
Pair 1	41%	46%	46%
Pair 2	46%	45%	
Pair 3	49%		

The data in this table are the percent-level of agreement results for the final ratings assigned by each testing pair for live ratings only.

Inter-Pair Reliability
Summary Results: Non-Parametric Analyses of Variance
English Pilot Study: Phase 1 Only

Table F-6. Inter-Pair Reliability as Measured by Non-Parametric Pearson Chi-Square
English Pilot Study: Phase 1 Only

	Pair 4			**Pair 3**			**Pair 2**		
	x^2	df	p	x^2	df	p	x^2	df	p
Pair 1	188.806	81	.0000**	137.428	72	.0000**	173.259	81	.0000**
Pair 2	184.317	81	.0000**	141.930	72	.0000**			
Pair 3	154.547	72	.0000**						

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings during the first half of the study only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. α = .05; *p<.05; **p<.01

Table F-7. Inter-Pair Reliability as Measured by Friedman Chi-Square of Ranks Test
English Pilot Study: Phase 1 Only

Testing Pair	**Median**	**Interquartile Range (IQR)**
Pair 1	2+	10.00
Pair 2	2+	20.00
Pair 3	2+	18.00
Pair 4	2+	12.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
8.5330	3	.0362*

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings during the first half of the study only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. α = .05; *p<.05; **p<.01

Table F-8. Inter-Pair Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, English Pilot Study: Phase 1 Only

	Pair 4		Pair 3		Pair 2	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
Pair 1	z = -1.7008 p = .0976	exact p = .0266*	z = -1.2356 p = .2316	exact p = .2632	z = -0.1230 p = .9258	exact p = 1.0000
Pair 2	z = -2.3185 p = .0171*	exact p = .0127*	z = -0.9712 p = .3428	exact p = .5034		
Pair 3	z = -0.9900 p = .3231	exact p = .3323				

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings during the first half of the study only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.
*α = .05;   *p<.05; **p<.01*

Table F-9. Inter-Pair Reliability as Measured by Kendall Tau-b Correlation Formula English Pilot Study: Phase 1 Only

	Pair 4	**Pair 3**	**Pair 2**
Pair 1	863	852	.881
Pair 2	910	871	
Pair 3	875		

The data in this table are the results of comparing all of the final negotiated ratings assigned by each testing pair for live ratings during the first half of the study only.

Table F-10. Inter-Pair Reliability as Measured by Percent Level of Agreement English Pilot Study: Phase 1 Only

	Pair 4	**Pair 3**	**Pair 2**
Pair 1	48%	46%	51%
Pair 2	58%	47%	
Pair 3	55%		

Th data in this table are the percent-level of agreement results for the final ratings assigned by each testing pair for live ratings during the first half of the study only.

Inter-Pair Reliability
Summary Results: Non-Parametric Analyses of Variance
English Pilot Study: Phase 2 Only

Table F-11. Inter-Pair Reliability as Measured by Non-Parametric Pearson Chi-Square English Pilot Study: Phase 2 Only

Testing Pair	**Pair 4**			**Pair 3**			**Pair 2**		
	x^2	df	p	x^2	df	p	x^2	df	p
Pair 1	112.379	42	.0000**	90.310	42	.0001**	59.962	30	.0007**
Pair 2	82.467	42	.0025**	76.374	42	.0014**			
Pair 3	85.241	49	.0013**						

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings during the first half of the study only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables.

*$\alpha = .05$; *p<.05; **p<.01*

Table F-12. Inter-Pair Reliability as Measured by Friedman Chi-Square of Ranks Test English Pilot Study: Phase 2 Only

	Median	**Interquartile Range (IQR)**
Pair 1	3	10.00
Pair 2	3	10.00
Pair 3	3	12.00
Pair 4	2+	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
15.6264	3	.0006**

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings during the first half of the study only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables.

*$\alpha = .05$; *p<.05; **p<.01*

Table F-13. Inter-Pair Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, English Pilot Study: Phase 2 Only

	Pair 4		**Pair 3**		**Pair 2**	
Test	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
Pair 1	z = -3.8138 p = .0001**	exact p = .0000**	z = -0.3828 p = .7282	exact p = 1.0000	z = -0.8420 p = .4386	exact p = .4807
Pair 2	z = -1.4975 p = .1434	exact p = .0784	z = -0.7077 p = .5005	exact p = 1.0000		
Pair 3	z = -2.3541 p = .0178*	exact p = .0075**				

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings during the second half of the study only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, those values are reported as **exact**. *Two-tailed probability value (p) is reported.*

$\alpha = .05$; $*p < .05$; $**p < .01$

Table F-14. Inter-Pair Reliability as Measured by Kendall Tau-b Correlation Formula English Pilot Study: Phase 2 Only

	Pair 4	**Pair 3**	**Pair 2**
Pair 1	.866	.663	.718
Pair 2	.690	.631	
Pair 3	.679		

The data in this table are the results of comparing all of the English final negotiated ratings assigned by each testing pair for live ratings during the second half of the study only.

Table F-15. Inter-Pair Reliability as Measured by Percent Level of Agreement English Pilot Study: Phase 2 Only

	Pair 4	**Pair 3**	**Pair 2**
Pair 1	33%	45%	40%
Pair 2	30%	42%	
Pair 3	42%		

The data in this table are the percent-level of agreement results for the final ratings assigned by each testing pair for live ratings during the second half of the study only.

Inter-Rater Reliability
Summary Results: Non-Parametric Correlations Between Individual Final Ratings Within Testing Pairs
English Pilot Study: Overall Study

Table F-16. Inter-Rater Reliability as Measured by Kendall Tau-b Correlation Formula English Pilot Study: Overall Study

Pair 1	Pair 2	Pair 3	Pair 4
n=75	n=75	n=74	n=75
.886	.845	.876	.966

The data in this table are the results of comparing the English individual tester ratings assigned by the testers in each pair for live ratings only.

Table F-17. Inter-Rater Reliability as Measured by Kendall Tau-b Correlation Formula English Pilot Study: Phase 1 Only

Pair 1	Pair 2	Pair 3	Pair 4
n=41	n=41	n=40	n=41
.921	.892	.864	.984

The data in this table are the results of comparing the English individual tester ratings assigned by the testers in each pair for the first half of the study for live ratings only.

Table F-18. Inter-Rater Reliability as Measured by Kendall Tau-b Correlation Formula English Pilot Study: Phase 2 Only

Pair 1	Pair 2	Pair 3	Pair 4
n=34	n=34	n=34	n=34
.798	.738	.872	.936

The data in this table are the results of comparing the English individual tester ratings assigned by the testers in each pair for the second half of the study for live ratings only.

Table F-19. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Kendall Tau-b Correlation Formula, Taped Ratings Only, English Pilot Study: Overall Study

All Pairs	Pair 1	Pair 2	Pair 3	Pair 4
n=96	n=24	n=24	n=21	n=24
.920	.829	.925	.881	1.000

The data in this table are the results of comparing the English individual tester ratings for taped ratings only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
English Pilot Study: Overall Study

Table F-20. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square
English Pilot Study: Overall Study

Test Order	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	230.402	81	.0000**	203.454	81	.0000**	251.868	81	.0000**
Second	236.117	81	.0000**	258.331	81	.0000**			
Third	250.912	81	.0000**						

The data in this table are the results of comparing all of the English final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables.
*α = .05; *p<.05; **p<.01*

Table F-21. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test
English Pilot Study: Overall Study

Test Order	Median	Interquartile Range (IQR)
First	3	18.00
Second	2+	18.00
Third	2+	18.00
Fourth	3	18.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
1.0888	3	.7828

The data in this table are the results of comparing all of the English final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables.
*α = .05; *p<.05; **p<.01*

BEST COPY AVAILABLE

Table F-22. Test Order Effects as Measured by Kendall Tau-b Correlation Formula
English Pilot Study: Overall Study

Test Order	Fourth	Third	Second
First	.802	.800	.821
Second	.770	.806	
Third	.756		

The data in this table are the results of comparing the English final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
English Pilot Study: Phase 1

Table F-23. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square
English Pilot Study: Phase 1 Only

Test Order	**Fourth**			**Third**			**Second**		
	x^2	df	p	x^2	df	p	x^2	df	p
First	144.048	72	.0000**	149.692	72	.0000**	154.548	72	.0000**
Second	182.227	81	.0000**	142.027	81	.0001**			
Third	167.685	81	.0000**						

The data in this table are the results of comparing all of the English final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only during the first half of the study. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. $\alpha = .05$; *p<.05; **p<.01

Table F-24. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test
English Pilot Study: Phase 1 Only

Test Order	**Median**	**Interquartile Range (IQR)**
First	2+	18.00
Second	2+	19.00
Third	2+	20.00
Fourth	2+	20.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
0.4387	3	.9343

The data in this table are the results of comparing all of the English final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only during the first half of the study. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. $\alpha = .05$; *p<.05; **p<.01

Table F-25. Test Order Effects as Measured by Kendall Tau-b Correlation Formula
English Pilot Study: Phase 1 Only

Test Order	Fourth	Third	Second
First	.871	.856	.874
Second	.845	.845	
Third	.813		

The data in this table are the results of comparing the English final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only during the first half of the study.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
English Pilot Study: Phase 2 Only

Table F-26. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square
English Pilot Study: Phase 2 Only

Test Order	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	81.590	42	.0005**	102.500	56	.0009**	76.541	42	.0003**
Second	63.433	30	.0000**	75.911	42	.0002**			
Third	107.812	48	.0001**						

The data in this table are the results of comparing the English final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only during the second half of the study. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. $\alpha = .05$; $*p<.05$; $**p<.01$

Table F-27. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test
English Pilot Study: Phase 2 Only

Test Order	Median	Interquartile Range (IQR)
First	3	10.00
Second	3	12.00
Third	3	12.00
Fourth	3	10.00
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
1.0879	3	.7843

The data in this table are the results of comparing the English final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only during the second half of the study. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. $\alpha = .05$; $*p<.05$; $**p<.01$

Table F-28. Test Order Effects as Measured by Kendall Tau-b Correlation Formula
English Pilot Study: Phase 2 Only

Test Order	Fourth	Third	Second
First	.643	.698	.727
Second	.587	.697	
Third	.609		

The data in this table are the results of comparing the English final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings only during the second half of the study.

Effects on Reliability Caused by Time of Administration
Summary Results: Non-Parametric Analyses of Variance
English Pilot Study: Overall Study

Table F-29. Test Slot Effects as Measured by Non-Parametric Pearson Chi-Square
English Pilot Study: Overall Study

Test Slot	**2:30 p.m.**			**1:00 p.m.**			**10:30 a.m.**		
	x^2	df	p	x^2	df	p	x^2	df	p
9:00 a.m.	175.336	81	.0000**	166.146	72	.0000**	204.757	81	.0000**
10:30 a.m.	184.840	81	.0000**	214.984	81	.0000**			
1:00 p.m.	206.026	81	.0000**						
Morning tests compared to afternoon tests							365.193	90	.0000**

The data in this table are the results of comparing the English final negotiated ratings assigned by all testing pairs in the tests administered in each testing slot for each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables.
*α = .05; *p<.05; **p<.01*

Table F-30. Test Slot Effects as Measured by Friedman Chi-Square of Ranks Test
English Pilot Study: Overall Study

Test Slot	**Median**	**Interquartile Range (IQR)**
9:00 a.m.	2+	18.00
10:30 a.m.	2+	18.00
1:00 p.m.	2+	10.00
2:30 p.m.	3	18.00
a.m. only	2	12.00
p.m. only	2	12.00
Friedman Two-way Anova Chi-Square of Ranks (comparing all four testing slots)		
x^2	**df**	**2-tailed p value**
7.6415	3	.0519
Friedman Two-way Anova Chi-Square of Ranks (comparing all am tests to all pm tests)		
x^2	**df**	**2-tailed p value**
1.2308	1	.3275

The data in this table are the results of comparing the English final negotiated ratings assigned by all testing pairs in the tests administered in each testing slot for each examinee for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. A.M. only *results combine the tests administered at 9:00 and 10:30, while p.m.* only *results combine the tests administered at 1:00 and 2:30.* α = .05; *p<.05; **p<.01

Table F-31. Test Slot Effects as Measured by Kendall Tau-b Correlation Formula
English Pilot Study: Overall Study

Test Slot	2:30 p.m.	1:00 p.m.	10:30 a.m.
9:00 a.m.	.832	.844	.863
10:30 a.m.	.823	.819	
1:00 p.m.	.794		
Morning Tests Compared to Afternoon Tests			.832

The data in this table are the results of comparing the English final negotiated ratings assigned by all testing pairs in the tests administered in each testing slot for each examinee for live ratings only.

Inter-Pair Reliability for Taped Ratings Only
Summary Results: Non-Parametric Analyses of Variance
English Pilot Study: Overall Study

Table F-32. Inter-Pair Reliability for Taped Ratings as Measured by Non-Parametric Pearson Chi-Square, English Pilot Study: Overall Study

Set of Testing Pairs	n	x^2	df	2-tailed p value
All Live Rating and All Taped Rating	84	404.060	100	.0000**
Pair 2 Live Rating and Pair 3 Taped Rating	6	12.000	12	1.0000
Pair 2 Live Rating and Pair 4 Taped Rating	5	15.000	9	.0997
Pair 2 Live Rating and Pair 5 Taped Rating	8	6.666	6	1.0000
Pair 3 Live Rating and Pair 2 Taped Rating	6	13.500	12	.9348
Pair 3 Live Rating and Pair 4 Taped Rating	7	15.750	16	1.0000
Pair 3 Live Rating and Pair 5 Taped Rating	8	17.777	12	.1318
Pair 4 Live Rating and Pair 2 Taped Rating	5	10.000	6	.2007
Pair 4 Live Rating and Pair 3 Taped Rating	7	22.750	20	.8521
Pair 4 Live Rating and Pair 4 Taped Rating	8	40.000	36	1.0000
Pair 5 Live Rating and Pair 2 Taped Rating	8	11.500	6	.0546
Pair 5 Live Rating and Pair 3 Taped Rating	8	20.666	16	.2564
Pair 5 Live Rating and Pair 4 Taped Rating	8	40.000	30	.0746

The data in this table are the results of comparing the English final negotiated ratings assigned by all testing pairs for taped ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables.
*$\alpha = .05$; *$p<.05$; **$p<.01$*

144

Table F-33. Inter-Pair Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, English Pilot Study: Overall

Set of Testing Pairs	n	Wilcoxon	Sign Test
All Live Rating and All Taped Rating	84	z = -1.9184 p = .0624	z = -1.0321 p = .3122
Pair 2 Live Rating and Pair 3 Taped Rating	6	z = -1.0690 p = .5039	exact p = 1.0000
Pair 2 Live Rating and Pair 4 Taped Rating	5	z = -1.4142 p = .4996	exact p = .5000
Pair 2 Live Rating and Pair 5 Taped Rating	8	z = -1.1862 p = .3825	exact p = .2188
Pair 3 Live Rating and Pair 2 Taped Rating	6	z = -0.7365 p = .6245	exact p = 1.0000
Pair 3 Live Rating and Pair 4 Taped Rating	7	z = -0.4121 p = .8091	exact p = 1.0000
Pair 3 Live Rating and Pair 5 Taped Rating	8	z = -0.7071 p = .7506	exact p = 1.0000
Pair 4 Live Rating and Pair 2 Taped Rating	5	z = -0.4472 p = 1.0000	exact p = 1.0000
Pair 4 Live Rating and Pair 3 Taped Rating	7	z = -1.6330 p = .2516	exact p = .2500
Pair 4 Live Rating and Pair 4 Taped Rating	8	z = -1.6330 p = .2453	exact p = .2500
Pair 5 Live Rating and Pair 2 Taped Rating	8	z = -0.7071 p = .7470	exact p = .3750
Pair 5 Live Rating and Pair 3 Taped Rating	8	z = -0.7428 p = .5018	exact p = 1.0000
Pair 5 Live Rating and Pair 4 Taped Rating	8	z = -1.8411 p = .1222	exact p = .1250

The data in this table are the results of comparing all of the English final negotiated ratings assigned by all testing pairs for taped ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables.

$\alpha = .05$; $^{*}p<.05$; $^{**}p<.01$

Table F-34. Inter-Pair Reliability for Taped Ratings as Measured by Kendall Tau-b Correlation Formula, English Pilot Study: Overall Study

Set of Testing Pairs	n	Correlations
All Live Rating and All Taped Rating	84	.805
Pair 2 Live Rating and Pair 3 Taped Rating	6	.889
Pair 2 Live Rating and Pair 4 Taped Rating	5	1.000
Pair 2 Live Rating and Pair 5 Taped Rating	8	.623
Pair 3 Live Rating and Pair 2 Taped Rating	6	.889
Pair 3 Live Rating and Pair 4 Taped Rating	7	.789
Pair 3 Live Rating and Pair 5 Taped Rating	8	.653
Pair 4 Live Rating and Pair 2 Taped Rating	5	.943
Pair 4 Live Rating and Pair 3 Taped Rating	7	.923
Pair 4 Live Rating and Pair 4 Taped Rating	8	.963
Pair 5 Live Rating and Pair 2 Taped Rating	8	.830
Pair 5 Live Rating and Pair 3 Taped Rating	8	.898
Pair 5 Live Rating and Pair 4 Taped Rating	8	.906

The data in this table are the results of comparing the English final negotiated ratings assigned by all testing pairs for taped ratings only.

Table F-35. Inter-Rater Reliability for Taped Ratings as Measured by Kendall Tau-b Correlation Formula , English Pilot Study: Overall Study

Set of Testing Pairs	n	Correlation
All Live Rating and All Taped Rating	96	.920
All Pair 1 Taped Ratings	24	.829
All Pair 2 Taped Ratings	24	.925
All Pair 3 Taped Ratings	21	.881
All Pair 4 Taped Ratings	24	1.000

The data in this table are the results of comparing the English individual final ratings assigned by each testing pair for taped ratings only.

Appendix G. Crosstab Charts for SPT Spanish Study

Crosstabulations
Interagency Reliability for Live Ratings
Overall Study
(SPT Spanish, 1994-95)

Chart G-1 Comparison of CIA and DLI

CIA \ DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	**1**											1
8	1	**1**										2
10		1	**3**									4
18			4	**9**	1							14
20			1	7	**9**	4	1					22
28				1	4	**10**	1					16
30					2	6	**6**					14
38						4	4	**4**	2	1		15
40						1	3	3	**4**	1		12
48						1		1	1			3
50								2			**2**	4
Totals		2	2	8	17	16	26	15	10	7	2	2	107

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-2. Comparison of CIA and FBI

FBI

CIA

	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10			3	1								4
18			1	7	3							11
20				1	13	6	2	1				23
28					7	8	1					16
30					2	8	3	1				14
38					1	4	4	6				15
40							4	7	1	1		13
48						1		1				2
50							1	1		1	1	4
Totals		1	2	4	9	26	27	15	17	1	2	1	105

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-3. Comparison of CIA and FSI

CIA \ FSI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10		1	2	1								4
18		1	2	10	1							14
20				5	16	3						24
28				3	8	5						16
30				2	4	6	3					15
38					2	6	8					16
40						1	6	4	1			12
48						1			1		1	3
50						1	1				2	4
Totals	1	4	4	21	31	23	18	4	2		3	111

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-4. Comparison of DLI and FBI

DLI \ FBI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1	1										2
8		1	1									2
10			2	6	1							9
18			1	4	9	2						16
20					7	6	1					14
28					10	14	5	1				30
30						5	6	5				16
38							3	6	1			10
40							2	3		2		7
48							1	2				3
50							1			1	1	3
Totals	1	2	4	10	27	27	19	17	1	3	1	112

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-5. Comparison of DLI and FSI

FSI

DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1	1										2
8		2										2
10		1	5	3								9
18				10	8							18
20				8	8	1						17
28				1	15	13	2					31
30					1	6	8	1				16
38					1	4	3	2			1	11
40							4	2	2			8
48							1	1				2
50										1	2	3
Totals	1	4	5	22	33	24	18	6	2	1	3	119

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-6. Comparison of FBI and FSI

FSI

FBI

	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10		1	1	2								4
18		1	4	5								10
20				8	16	4						28
28				4	13	10	2					29
30					5	4	10			1		20
38						5	6	4	1		1	17
40								1				1
48							1	1			1	3
50											1	1
Totals	1	4	5	19	34	23	19	6	1	1	3	116

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Interagency Reliability for Live Ratings
Team 1 Only
(SPT Spanish, 1994-95)

Chart G-7. Comparison of CIA and DLI

DLI

		0	8	10	18	20	28	30	38	40	48	50	Totals
CIA	0	1											1
	8	1	1										2
	10		1										1
	18			3	6								9
	20				5	4	3	1					13
	28					2	4						6
	30					2	3	2					7
	38						2	2	2	1	1		8
	40						1	1	1	3			6
	48						1		1	1			3
	50											2	2
Totals		2	2	3	11	8	14	6	4	5	1	2	58

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-8. Comparison of CIA and FBI

FBI

CIA

	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10			1									1
18			1	4	2							7
20					6	5	1	1				13
28					2	4						6
30						5	1					6
38						2	2	4				8
40							1	3	1	1		6
48						1		1				2
50										1	1	2
Totals	1	2	2	4	10	17	5	9	1	2	1	54

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-9. Comparison of CIA and FSI

FSI

		0	8	10	18	20	28	30	38	40	48	50	Totals
CIA	0	1											1
	8		2										2
	10		1										1
	18		1	1	6	1							9
	20				3	7	3						13
	28				2	2	2						6
	30				2	1	3	1					7
	38					2	2	4					8
	40							3	2	1			6
	48						1			1		1	3
	50											2	2
Totals		1	4	1	13	13	11	8	2	2		3	58

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-10. Comparison of DLI and FBI

DLI \ FBI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1	1										2
8		1	1									2
10				3	1							4
18			1	2	5	2						10
20					3	5						8
28					2	8	5	1				16
30						3		3				6
38								3	1			4
40							1	2		1		4
48							1					1
50										1	1	2
Totals	1	2	2	5	11	18	7	9	1	2	1	59

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-11. Comparison of DLI and FSI

FSI

DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1	1										2
8		2										2
10		1	2	1								4
18				7	5							12
20				5	2	1						8
28				1	7	7	2					17
30					1	2	3					6
38						1	1	1			1	4
40							2	1	2			5
48							1					1
50											2	2
Totals	1	4	2	14	15	11	9	2	2		3	63

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-12. Comparison of FBI and FSI

FSI

FBI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10		1		1								2
18		1	2	2								5
20				5	5	1						11
28				4	7	5	2					18
30					3	1	3					7
38						3	3	1	1		1	9
40								1				1
48							1				1	2
50											1	1
Totals	1	4	2	12	15	10	9	2	1		3	59

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Interagency Reliability for Live Ratings
Team 2 Only
(SPT Spanish, 1994-95)

Chart G-13. Comparison of CIA and DLI

CIA \ DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			3									3
18			1	3	1							5
20			1	2	5	1						9
28				1	2	6	1					10
30						3	4					7
38						2	2	2	1			7
40							2	2	1	1		6
48												
50								2				2
Totals			5	6	8	12	9	6	2	1		49

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-14. Comparison of CIA and FBI

FBI

CIA

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2	1								3
18				3	1							4
20				1	7	1	1					10
28					5	4	1					10
30					2	3	2	1				8
38					1	2	2	2				7
40							3	4				7
48												
50							1	1				2
Totals			2	5	16	10	10	8				51

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-15. Comparison of CIA and FSI

CIA \ FSI	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2	1								3
18			1	4								5
20				2	9							11
28				1	6	3						10
30					3	3	2					8
38						4	4					8
40						1	3	2				6
48												
50						1	1					2
Totals			3	8	18	12	10	2				53

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-16. Comparison of DLI and FBI

FBI

DLI

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2	3								5
18				2	4							6
20					4	1	1					6
28					8	6						14
30						2	6	2				10
38							3	3				6
40							1	1		1		3
48								2				2
50							1					1
Totals			2	5	16	9	12	8		1		53

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-17. Comparison of DLI and FSI

FSI

DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			3	2								5
18				3	3							6
20				3	6							9
28					8	6						14
30						4	5	1				10
38					1	3	2	1				7
40							2	1				3
48								1				1
50										1		1
Totals			3	8	18	13	9	4		1		56

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-18. Comparison of FBI and FSI

FSI

FBI

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			1	1								2
18			2	3								5
20				3	11	3						17
28					6	5						11
30					2	3	7			1		13
38						2	3	3				8
40												
48								1				1
50												
Totals			3	7	19	13	10	4		1		57

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Interagency Reliability for Live Ratings
Phase 1 Only
(SPT Spanish, 1994-95)

Chart G-19. Comparison of CIA and DLI

CIA \ DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8	1	1										2
10			3									3
18			2	4	1							7
20				5	3	4						12
28					2	4	1					7
30						2	3					5
38						2	2	2		1		7
40						1		1	1			3
48												
50												
Totals	2	1	5	9	6	13	6	3	1	1		47

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-20. Comparison of CIA and FBI

FBI

CIA	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10			2	1								3
18			1	4								5
20					8	4	2					14
28					2	4	1					7
30						2	1	1				4
38					1	2	3	1				7
40							1	2				3
48												
50												
Totals	1	2	3	5	11	12	8	4	0	0	0	46

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-21. Comparison of CIA and FSI

FSI

CIA	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10			2	1								3
18		1	1	5								7
20				2	10	2						14
28				1	4	2						7
30						4	1					5
38					2	2	4					8
40							1	1	1			3
48												
50												
Totals	1	3	3	9	16	10	6	1	1			50

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-22. Comparison of DLI and FBI

FBI

DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1	1										2
8		1										1
10			2	3								5
18			1	2	5	1						9
20					2	2	1					5
28					4	7	3					14
30						2	3	1				6
38								2				2
40								1		1		2
48							1	1				2
50												
Totals	1	2	3	5	11	12	8	5		1		48

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-23. Comparison of DLI and FSI

FSI

DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1	1										2
8		1										1
10		1	3	1								5
18				6	4							10
20				2	4	1						7
28				1	7	5	2					15
30					1	3	2					6
38						1	1	1				3
40								1	1			2
48							1	1				2
50												
Totals	1	3	3	10	16	10	6	3	1			53

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-24. Comparison of FBI and FSI

FSI

FBI

	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10			1	2								3
18		1	2	2								5
20				4	6	2						12
28					9	3	1					13
30					2	3	4					9
38						1	1	2	1			5
40												
48								1				1
50												
Totals	1	3	3	8	17	9	6	3	1			51

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Interagency Reliability for Live Ratings
Phase 2 Only
(SPT Spanish, 1994-95)

Chart G-25. Comparison of CIA and DLI

DLI

CIA	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10		1										1
18			2	**5**								7
20			1	2	**6**		1					10
28				1	2	**6**						9
30					2	4	**3**					9
38						2	2	**2**	2			8
40							3	2	**3**			8
48						1		1	1			3
50								2			**2**	4
Totals		1	3	8	10	13	9	7	6		2	59

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-26. Comparison of CIA and FBI

FBI

CIA	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			**1**									1
18				**3**	3							6
20				1	**5**	2		1				9
28					5	**4**						9
30					2	6	**2**					10
38						2	1	**5**				8
40							3	4	**1**	1		9
48						1		1				2
50							1	1		1	**1**	4
Totals			1	4	15	15	7	12	1	2	1	58

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-27. Comparison of CIA and FSI

CIA \ FSI	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10		1										1
18			1	**5**	1							7
20				3	**6**	1						10
28				2	4	**3**						9
30				2	4	2	**2**					10
38						4	4					8
40						1	5	3				9
48						1			1		1	3
50						1	1				**2**	4
Totals		1	1	12	15	13	12	3	1		3	61

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-28. Comparison of DLI and FBI

FBI

DLI

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8			1									1
10				3	1							4
18				**2**	4	1						7
20					**5**	4						9
28					6	**7**	2	1				16
30						3	**3**	4				10
38							3	**4**	1			8
40							2	2		1		5
48												
50							1			1	**1**	3
Totals				1	5	16	15	11	11	1	2	1	63

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-29. Comparison of DLI and FSI

FSI

DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8		**1**										1
10			**2**	2								4
18				**4**	4							8
20				6	**4**							10
28					8	**8**						16
30						3	**6**	1				10
38					1	3	2	**1**			1	8
40							4	1	**1**			6
48												
50										1	**2**	3
Totals		**1**	**2**	**12**	**17**	**14**	**12**	**3**	**1**	**1**	**3**	**66**

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-30. Comparison of FBI and FSI

FSI

FBI

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10		1										1
18			2	3								5
20				4	10	2						16
28				4	4	7	1					16
30					3	1	6			1		11
38						4	5	2			1	12
40								1				1
48							1				1	2
50											1	1
Totals		1	2	11	17	14	13	3		1	3	65

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Interagency Reliability for Taped Ratings
(SPT Spanish, 1994-95)

Chart G-31. Comparison of CIA and DLI

DLI

CIA	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2									2
18			2	5								7
20			1	2	3	3	1					10
28				1	4	3	2	1				11
30						2	3	1				6
38							2	2				4
40								2				2
48											1	1
50												
Totals			5	8	7	8	8	6			1	43

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-32. Comparison of CIA and FBI

FBI

CIA	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			1	1								2
18				4	3							7
20				3	2	4	1					10
28					4	5	1	1				11
30					1	2	2	1				6
38						1	2		1			4
40								1			1	2
48									1			1
50												
Totals			1	8	10	12	6	3	2		1	43

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-33. Comparison of CIA and FSI

FSI

CIA

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			1	1								2
18			2	4	1							7
20				5	3	1		1				10
28				2	4	3	2					11
30						5	1					6
38						3	1					4
40							1		1			2
48								1				1
50												
Totals			3	12	8	12	5	2	1			43

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-34. Comparison of DLI and FBI

FBI

DLI	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			1	4								5
18				4	4							8
20					4	4						8
28					1	7	1	1				10
30					1	3	3	1				8
38							3	2	1			6
40									2	1		3
48												
50												
Totals			1	8	10	14	7	4	3	1		48

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

BEST COPY AVAILABLE

Chart G-35. Comparison of DLI and FSI

FSI

DLI

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			1	3	1							5
18			2	6								8
20				2	3	2	1					8
28				1	4	4	1					10
30					1	4	2	1				8
38						3	2		1			6
40								1	2			3
48												
50												
Totals			3	12	9	13	6	2	3			48

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-36. Comparison of FBI and FSI

FSI

		0	8	10	18	20	28	30	38	40	48	50	Totals
FBI	0												
	8												
	10				1								1
	18			2	5	1							8
	20			1	3	2	3	1					10
	28				3	5	5	1					14
	30						4	1	1	1			7
	38					1		3					4
	40						1		1	1			3
	48									1			1
	50												
Totals				3	12	9	13	6	2	3			48

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Intra-agency Reliability for Live Ratings vs. Taped Ratings
(SPT Spanish, 1994-95)

Chart G-37. CIA Comparisons

CIA — *Taped* (columns), *Live* (rows)

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10												
18			2	**5**								7
20				1	**5**	3						9
28					3	**3**	1					7
30						2	**2**					4
38					1	2	1	**2**				6
40							2	2	**2**			6
48												
50												
Totals			2	6	9	10	6	4	2			39

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-38. DLI Comparisons

DLI *Taped*

Live

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			**1**	2								3
18			3	**2**	2							7
20				3	**3**	2						8
28				1	2	**7**	1	1				12
30					1		**4**	2				7
38							2	**2**				4
40						1			**1**			2
48							1		1			2
50									1			1
Totals			4	8	8	10	8	5	3			46

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-39. FBI Comparisons

FBI *Taped*

Live	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10												
18			1	**4**								5
20				2	**4**	5	1					12
28				2	4	**6**	1					13
30						2	**1**	2		1		6
38							2	**1**	3			6
40							1					1
48								1				1
50												
Totals			1	8	8	13	6	4	3	1		44

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-40. FSI Comparisons

FSI *Taped*

Live

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			**1**	1								2
18			2	**7**	3							12
20				4	**4**	3	1					12
28					2	**5**	1					8
30						5	**2**	1				8
38							2	**1**	1	0		4
40									**1**			1
48									1			1
50												
Totals			3	12	9	13	6	2	3			48

ILR ratings are represented by the following codes:

ILR Rati	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Live Ratings
Overall Study
(SPT Spanish, 1994-95)

Chart G-41. CIA Comparisons

CIA

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	1	1										2
	8		1										1
	10			3	1								4
	18			1	12	3							16
	20				2	19	4	1					26
	28					3	14	2					19
	30						1	12	2				15
	38							1	15	4			20
	40									12	1		13
	48										3	2	5
	50											4	4
Totals		1	2	4	15	25	19	16	17	16	4	6	125

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-42. DLI Comparisons

DLI

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	2											2
	8		2										2
	10			7	3								10
	18				16	2							18
	20					16	2						18
	28					1	29	2					32
	30						1	14	3				18
	38							1	9	2			12
	40									6			6
	48										4		4
	50											3	3
Totals		2	2	7	19	19	32	17	12	8	4	3	125

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-43. FBI Comparisons

FBI

Tester 2

Tester 1

	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10			4									4
18				9	6							15
20				2	24	3	1					30
28					2	24	1					27
30						3	16	2				21
38						2	1	11	1			15
40								5				5
48									1	3		4
50											1	1
Totals	1	2	4	11	32	32	19	18	2	3	1	125

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-44. FSI Comparisons

FSI

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	1											1
	8		4										4
	10			5									5
	18				23								23
	20					35							35
	28						24						24
	30							20					20
	38							1	6				7
	40									2			2
	48										1		1
	50											3	3
Totals		1	4	5	23	35	24	21	6	2	1	3	125

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Live Ratings
Team 1 Only
(SPT Spanish, 1994-95)

Chart G-45. CIA Comparisons

CIA *Tester 2*

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1	1										2
8		1										1
10												
18			1	8	2							11
20				2	9	2	1					14
28					3	4	1					8
30						1	4	2				7
38							1	7	1			9
40									6			6
48										3		3
50											2	2
Totals	1	2	1	10	14	7	7	9	7	3	2	63

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-46. DLI Comparisons

DLI

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	2											2
	8		2										2
	10			4									4
	18				12								12
	20					8							8
	28						17						17
	30							6					6
	38								4				4
	40									4			4
	48										2		2
	50											2	2
Totals		2	2	4	12	8	17	6	4	4	2	2	63

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-47. FBI Comparisons

FBI

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	1											1
	8		2										2
	10			2									2
	18				5	2							7
	20				2	8	2	1					13
	28					2	13						15
	30						3	5					8
	38						2	1	6				9
	40								2				2
	48									1	2		3
	50											1	1
Totals		1	2	2	7	12	20	7	8	1	2	1	63

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-48. FSI Comparisons

FSI *Tester 2*

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		4										4
10			2									2
18				14								14
20					15							15
28						11						11
30							9					9
38								2				2
40									2			2
48												
50											3	3
Totals	1	4	2	14	15	11	9	2	2		3	63

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Live Ratings
Team 2 Only
(SPT Spanish, 1994-95)

Chart G-49. CIA Comparisons

CIA *Tester 2*

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			3	1								4
18				4	1							5
20					10	2						12
28						10	1					11
30							8					8
38								8	3			11
40									6	1		7
48											2	2
50											2	2
Totals			3	5	11	12	9	8	9	1	4	62

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-50. DLI Comparisons

DLI

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8												
	10			3	3								6
	18				4	2							6
	20					8	2						10
	28					1	12	2					15
	30						1	8	3				12
	38							1	5	2			8
	40									2			2
	48										2		2
	50											1	1
Totals				3	7	11	15	11	8	4	2	1	62

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-51. FBI Comparisons

FBI

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8												
	10			2									2
	18				4	4							8
	20					16	1						17
	28						11	1					12
	30							11	2				13
	38								5	1			6
	40								3				3
	48										1		1
	50												
Totals				2	4	20	12	12	10	1	1		62

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-52. FSI Comparisons

FSI

Tester 2

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			3									3
18				9								9
20					20							20
28						13						13
30							11					11
38							1	4				5
40												0
48										1		1
50												
Totals			3	9	20	13	12	4		1		62

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Live Ratings
Phase 1 Only
(SPT Spanish, 1994-95)

Chart G-53. CIA Comparisons

CIA — *Tester 2* (columns), *Tester 1* (rows)

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	1	1										2
	8		1										1
	10			3	1								4
	18				5	1							6
	20				2	11	3	1					17
	28					2	6						8
	30							4	1				5
	38								7	2			9
	40									2	1		3
	48											1	1
	50												
Totals		1	2	3	8	14	9	5	8	4	1	1	56

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-54. DLI Comparisons

DLI *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	2											2
	8		1										1
	10			3	2								5
	18				9	2							11
	20					7	1						8
	28						15						15
	30						1	4	2				7
	38							1	1	2			4
	40												
	48										3		3
	50												
Totals		2	1	3	11	9	17	5	3	2	3		56

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-55. FBI Comparisons

FBI

Tester 2

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0	1											1
8		2										2
10			3									3
18				5	4							9
20				1	9	2	1					13
28					1	9	1					11
30						3	5	2				10
38						1	1	1	1			4
40								2				2
48										1		1
50												
Totals	1	2	3	6	14	15	8	5	1	1		56

ILR ratings are represented by the following codes

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-56. FSI Comparisons

FSI *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	1											1
	8		3										3
	10			3									3
	18				11								11
	20					17							17
	28						10						10
	30							7					7
	38								3				3
	40									1			1
	48												
	50												
Totals		1	3	3	11	17	10	7	3	1			56

ILR ratings are represented by the following codes.

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Live Ratings
Phase 2 Only
(SPT Spanish, 1994-95)

Chart G-57. CIA Comparisons

CIA

Tester 2

Tester 1

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10												
18			1	7	2							10
20					8	1						9
28					1	8	2					11
30						1	8	1				10
38							1	8	2			11
40									9			9
48										3	1	4
50											4	4
Totals			1	7	11	10	11	9	11	3	5	68

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-58. DLI Comparisons

DLI *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8		1										1
	10			4	1								5
	18				7								7
	20					9	1						10
	28					1	14	2					17
	30							10	1				11
	38								8				8
	40									6			6
	48												
	50											3	3
Totals			1	4	8	10	15	12	9	6		3	68

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-59. FBI Comparisons

FBI *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8												
	10			1									1
	18				4	2							6
	20				1	15	1						17
	28					1	15						16
	30							11					11
	38						1		9				10
	40								3				3
	48									1	2		3
	50											1	1
Totals				1	5	18	17	11	12	1	2	1	68

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-60. FSI Comparisons

FSI

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8		1										1
	10			2									2
	18				12								12
	20					18							18
	28						14						14
	30							13					13
	38								3				3
	40									1			1
	48										1		1
	50											3	3
Totals			1	2	12	18	14	13	3	1	1	3	68

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Taped Ratings
(SPT Spanish, 1994-95)

Chart G-61. CIA Comparisons

CIA *Tester 2*

Tester 1

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2									2
18				7								7
20				1	8	1	1					11
28					1	10						11
30						1	6					7
38								4	1			5
40									2			2
48										1		1
50												
Totals			2	8	9	12	7	4	3	1		46

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-62. DLI Comparisons

DLI

Tester 2

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			5									5
18				8								8
20					6	1						7
28					1	10						11
30							8	1				9
38								5				5
40									3			3
48												
50												
Totals			5	8	7	11	8	6	3			48

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-63. FBI Comparisons

FBI *Tester 2*

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10												
18			1	7	1							9
20					11							11
28					1	12		1		0		14
30						1	5			0		6
38								4	1	0		5
40								1	1	0		2
48									1	0		1
50										0		
Totals			1	7	13	13	5	6	3			48

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart G-64. FSI Comparisons

FSI

Tester 2

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8												
	10			3									3
	18				12								12
	20					9							9
	28						13						13
	30							5	1				6
	38								2				2
	40									3			3
	48												
	50												
Totals				3	12	9	13	5	3	3			48

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Appendix H. Crosstab Charts for SPT English Study

Crosstabulations
Inter-pair Reliability for Live Ratings
Overall Study
(SPT English, 1995)

Chart H-1. Comparison of Pair 1 and Pair 2

Pair 2

Pair 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8	1											1
10		1	3									4
18			1	4								5
20			1	3	5	1						10
28					1	8	4					13
30						3	6	7	2			18
38						1	1	2	1			5
40							1	4	1			6
48									2		1	3
50									1		3	4
Totals	1	1	5	7	6	13	12	13	7		4	69

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-2. Comparison of Pair 1 and Pair 3

Pair 3

Pair 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8		**1**										1
10		1	**3**									4
18			2	**1**	1							4
20			1	3	**5**							9
28					4	**6**	3					13
30					1	3	**6**	5	1	2		18
38							2	**1**	1	2		6
40						1		3	**2**			6
48								1		**2**		3
50											**4**	4
Totals		2	6	4	11	10	11	10	4	6	4	68

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-3. Comparison of Pair 1 and Pair 4

Pair 4

		0	8	10	18	20	28	30	38	40	48	50	Totals
Pair 1	0												
	8		1										1
	10			4									4
	18			3	2								5
	20			1	5	3	1						10
	28					4	9						13
	30						12	2	4				18
	38						1	1	2	1			5
	40						1	1	5				7
	48										3		3
	50									1		3	4
Totals			1	8	7	7	24	4	11	2	3	3	70

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-4. Comparison of Pair 2 and Pair 3

Pair 2 \ Pair 3	0	8	10	18	20	28	30	38	40	48	50	Totals
0		1										1
8		1										1
10			4	1								5
18			2	1	3							6
20				2	3	1						6
28					5	5	3					13
30						3	6		3			12
38						2		8		3		13
40							2	3		1	1	7
48												
50										2	3	5
Totals		2	6	4	11	11	11	11	3	6	4	69

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-5. Comparison of Pair 2 and Pair 4

		Pair 4											
		0	**8**	**10**	**18**	**20**	**28**	**30**	**38**	**40**	**48**	**50**	**Totals**
Pair 2	0		1										1
	8			1									1
	10			**5**									5
	18			2	**5**								7
	20				2	**3**	1						6
	28					4	**10**						14
	30						8	**1**	2	1			12
	38						5	3	**4**	1			13
	40						1		2	**1**	2		6
	48												
	50										2	**3**	5
Totals			1	8	7	7	25	4	8	3	4	3	70

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-6. Comparison of Pair 3 and Pair 4

Pair 4

Pair 3

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8		1	1									2
10			5	1								6
18			1	3								4
20				3	6	2						11
28					1	8	1	1				11
30						8	1	1				10
38						4	2	3	1	1		11
40						1		2	1			4
48						1		2		3		6
50									1		3	4
Totals		1	7	7	7	24	4	9	3	4	3	69

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-pair Reliability for Live Ratings
Phase 1 Only
(SPT English, 1995)

Chart H-7. Comparison of Pair 1 and Pair 2

Pair 2

Pair 1

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8	1											1
10		1	1									2
18			1	4								5
20				3	4	1						8
28						5	2					7
30						1	2	4	1			8
38						1						1
40								1	1			2
48											1	1
50									1		3	4
Totals	1	1	2	7	4	8	4	5	3		4	39

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-8. Comparison of Pair 1 and Pair 3

Pair 3

Pair 1

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8		1										1
10		1	1									2
18			2	1	1							4
20			1	1	5							7
28					3	2	2					7
30					1	1	2	3		1		8
38							1					1
40								2				2
48										1		1
50											4	4
Totals		2	4	2	10	3	5	5		2	4	37

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-9. Comparison of Pair 1 and Pair 4

Pair 4

Pair 1

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8		1										1
10			2									2
18			3	2								5
20				4	3	1						8
28					2	5						7
30						3	1	4				8
38						1		1				2
40						1		1				2
48										1		1
50									1		3	4
Totals		1	5	6	5	11	1	6	1	1	3	40

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-10. Comparison of Pair 2 and Pair 3

Pair 2 \ Pair 3	0	8	10	18	20	28	30	38	40	48	50	Totals
0		1										
8		1										1
10			2									2
18			2	1	3							6
20				1	3							4
28					4	2	2					8
30						1	3					4
38						1		3		1		5
40								2			1	3
48												
50										1	3	4
Totals		2	4	2	10	4	5	5		2	4	38

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-11. Comparison of Pair 2 and Pair 4

Pair 4

Pair 2	0	8	10	18	20	28	30	38	40	48	50	Totals
0		1										1
8			1									1
10			**2**									2
18			2	**5**								7
20				1	**3**							4
28					2	**7**						9
30						3		1				4
38						2	1	**2**				5
40								2	**1**			3
48												
50										1	**3**	4
Totals		**1**	**5**	**6**	**5**	**12**	**1**	**5**	**1**	**1**	**3**	**40**

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-12. Comparison of Pair 3 and Pair 4

Pair 4

		0	8	10	18	20	28	30	38	40	48	50	Totals
Pair 3	0												
	8		1	1									2
	10			3	1								4
	18				2								2
	20				3	5	2						10
	28						3	1					4
	30						4		1				5
	38						2		3				5
	40												
	48								1		1		2
	50									1		3	4
Totals			1	4	6	5	11	1	5	1	1	3	38

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-pair Reliability for Live Ratings
Phase 2 Only
(SPT English, 1995)

Chart H-13. Comparison of Pair 1 and Pair 2

Pair 2

Pair 1

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2									2
18												
20			1		1							2
28					1	3	2					6
30						2	4	3	1			10
38							1	2	1			4
40							1	3				4
48									2			2
50												
Totals			3		2	5	8	8	4			30

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-14. Comparison of Pair 1 and Pair 3

Pair 1 \ Pair 3	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2									2
18												
20				2								2
28					1	4	1					6
30						2	4	2	1	1		10
38							1	1	1	2		5
40						1		1	2			4
48								1		1		2
50												
Totals			2	2	1	7	6	5	4	4		31

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-15. Comparison of Pair 1 and Pair 4

Pair 4

		0	8	10	18	20	28	30	38	40	48	50	Totals
Pair 1	0												
	8												
	10			2									2
	18												
	20			1	1								2
	28					2	4						6
	30						9	1					10
	38							1	1	1			3
	40							1	4				5
	48										2		2
	50												
Totals				3	1	2	13	3	5	1	2		30

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-16. Comparison of Pair 2 and Pair 3

Pair 2 \ Pair 3	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2	1								3
18												
20				1		1						2
28					1	3	1					5
30						2	3		3			8
38						1		5		2		8
40							2	1		1		4
48												
50										1		1
Totals			2	2	1	7	6	6	3	4		31

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-17. Comparison of Pair 2 and Pair 4

Pair 2 \ Pair 4	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			3									3
18												
20				1		1						2
28					2	3						5
30						5	1	1	1			8
38						3	2	2	1			8
40						1				2		3
48												
50										1		1
Totals			3	1	2	13	3	3	2	3		30

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-18. Comparison of Pair 3 and Pair 4

Pair 3 \ Pair 4	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2									2
18			1	1								2
20					1							1
28					1	5		1				7
30						4	1					5
38						2	2		1	1		6
40						1		2	1			4
48						1		1		2		4
50												
Totals			3	1	2	13	3	4	2	3		31

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Live Ratings
Overall Study
(SPT English, 1995)

Chart H-19. Pair 1 Comparisons

Pair 1 *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8		1										1
	10			3	1								4
	18				2								2
	20				3	10							13
	28						10	1					11
	30					1	2	13	5				21
	38							2	2	2			6
	40							2	2	2	2		8
	48									1	4		5
	50											4	4
Totals			1	3	6	11	12	18	9	5	6	4	75

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-20. Pair 2 Comparisons

Pair 2 *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	1											1
	8		1										1
	10			3	3								6
	18				5	2							7
	20				1	3	2						6
	28					2	5	1					8
	30						6	6	3				15
	38							1	8	3			12
	40						1	2	5	1			9
	48									2			2
	50									3		5	8
Totals		1	1	3	9	7	14	10	16	9		5	75

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-21. Pair 3 Comparisons

Pair 3 *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8		1										1
	10			5									5
	18			1	1	2							4
	20			1	3	8		1					13
	28					2	6	4					12
	30						2	7	4	2			15
	38							1	6	3			10
	40								1	2	1		4
	48									1	6		7
	50											3	3
Totals			1	7	4	12	8	13	11	8	7	3	74

ILR ratings are represented by the following codes.

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-22. Pair 4 Comparisons

Pair 4 ***Tester 2***

Tester 1

	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8		1										1
10			7									7
18			1	7								8
20					7							7
28						23	2					25
30						2	3					5
38							1	11				12
40									2	1		3
48										4		4
50											3	3
Totals		1	8	7	7	25	6	11	2	5	3	75

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Live Ratings
Phase 1 Only
(SPT English 1995)

Chart H-23. Pair 1 Comparisons

Pair 1 *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8		1										1
	10			1	1								2
	18				2								2
	20				3	8							11
	28						6	1					7
	30						1	5	2				8
	38							2	1				3
	40										1		1
	48									1	1		2
	50											4	4
Totals			1	1	6	8	7	8	3	1	2	4	41

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-24. Pair 2 Comparisons

Pair 2 *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	1											1
	8		1										1
	10			1	2								3
	18				5	2							7
	20				1	2	2						5
	28					1	3						4
	30						3	3	1				7
	38								3	2			5
	40							1	1				2
	48									1			1
	50									1		4	5
Totals		1	1	1	8	5	8	4	5	4		4	41

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-25. Pair 3 Comparisons

Pair 3 ***Tester 2***

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0	1											1
	8		1										1
	10			4									4
	18					1							1
	20			1	3	7		1					12
	28					2	2	2					6
	30							4	2	1			7
	38							1	1	1			3
	40								1		1		2
	48										1		1
	50											3	3
Totals			1	5	3	10	2	8	4	2	2	3	40

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-26. Pair 4 Comparisons

Pair 4 *Tester 2*

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8		1										1
10			4									4
18			1	6								7
20					5							5
28						11	1					12
30							1					1
38								6				6
40										1		1
48										1		1
50											3	3
Totals		1	5	6	5	11	2	6		2	3	41

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Crosstabulations
Inter-rater Reliability for Live Ratings
Phase 2 Only
(SPT English, 1995)

Chart H-27. Pair 1 Comparisons

Pair 1 *Tester 2*

Tester 1	0	8	10	18	20	28	30	38	40	48	50	Totals
0												
8												
10			2									2
18												
20					2							2
28						4						4
30					1	1	8	3				13
38								1	2			3
40							2	2	2	1		7
48										3		3
50												
Totals			2		3	5	10	6	4	4		34

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-28. Pair 2 Comparisons

Pair 2 *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8												
	10			2	1								3
	18												
	20					1							1
	28					1	2	1					4
	30						3	3	2				8
	38							1	5	1			7
	40						1	1	4	1			7
	48									1			1
	50									2		1	3
Totals				2	1	2	6	6	11	5		1	34

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-29. Pair 3 Comparisons

Pair 3 *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8												
	10			1									1
	18			1	1	1							3
	20					1							1
	28						4	2					6
	30						2	3	2	1			8
	38								5	2			7
	40									2			2
	48									1	5		6
	50												
Totals				2	1	2	6	5	7	6	5		34

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50

Chart H-30. Pair 4 Comparisons

Pair 4 *Tester 2*

		0	8	10	18	20	28	30	38	40	48	50	Totals
Tester 1	0												
	8												
	10			3									3
	18				1								1
	20					2							2
	28						12	1					13
	30						2	2					4
	38							1	5				6
	40									2			2
	48										3		
	50												
Totals				3	1	2	14	4	5	2	3		34

ILR ratings are represented by the following codes:

ILR Rating	0	0+	1	1+	2	2+	3	3+	4	4+	5
Code	0	8	10	18	20	28	30	38	30	48	50